Topics in Speech and Audio Processing in Adverse Environments

Editors:

Eberhard Hänsler and Gerhard Schmidt

(The final design of the title page is done by Springer.)

List of Contributors

R. Aichner
Microsoft Corporation
Redmond, WA, USA

H. Buchner
Deutsche Telekom Laboratories
Germany

M. Buck
Harman/Becker
Germany

M. Christoph
Harman/Becker
Germany

I. Cohen
Israel Institute of Technology
Israel

S. Gannot
Bar-Ilan University
Israel

H. W. Gierlich
HEAD acoustics
Germany

E. Habets
Bar-Ilan University
Israel

E. Hänsler
Darmstadt University of Technology
Germany

T. Haulick
Harman/Becker
Germany

S. Haykin
McMaster University
Canada

U. Heute
University of Kiel
Germany

O. Hoshuyama
NEC Corporation
Japan

B. Iser
Harman/Becker
Germany

W. Kellermann
University Erlangen-Nuremberg
Germany

F. Kettler
HEAD acoustics
Germany

M. Krini
Harman/Becker
Germany

H. W. Löllmann
RWTH Aachen University
Germany

H.-J. Pfleiderer
University Ulm
Germany

H. Puder
Siemens Audiological Engineering Group
Germany

R. Rabenstein
University Erlangen-Nuremberg
Germany

N. Roman
Ohio State University at Lima
Lima, USA

J. Scheuing
University of Stuttgart
Germany

G. Schmidt
Harman/Becker
Germany

A. Sehr
University Erlangen-Nuremberg
Germany

S. Spors
Deutsche Telekom Laboratories
Germany

A. Sugiyama
NEC Corporation
Japan

P. Vary
RWTH Aachen University
Germany

D. Wang
Ohio State University
Columbus, USA

K. Wiklund
McMaster University
Canada

B. Yang
University of Stuttgart
Germany

Contents

Part VI Selected Applications

Abbreviations and Acronyms

AA-LP	Anti-aliasing lowpass
AC3	Adaptive transform coder 3
ACR	Absolute-category rating
ADC	Analog-to-digital converter
ADPCM	Adaptive differential pulse-code modulation
AEC	Acoustic echo cancellation / caneller
AIR	Acoustic impulse response
ALP	Adaptive lattice predictor
ANC	Active noise control or adaptive noise canceller
AP	Affine projection
APC	Adaptive predictive coding
AR	Auto-regressive
AS	Analysis synthesis
ASA	Auditory scene analysis
ASR	Automatic speech recognition
ATC	Adaptive transform coding
BSA	Bark-spectral approximation
BSD	Bark-spectral distance
BSS	Blind source separation
CAN	Controller area network
CASA	Computational auditory scene analysis
CB	Critical band
CCR	Comparison-category rating
CD	Cepstral distance
CELP	Code-excited linear predictive (coding)
CMN	Cepstral mean normalization
CMS	Cepstral mean subtraction
CPU	Central processing unit
DAC	Digital-to-analog converter
DAM	Diagnostic acceptability measure
DCR	Degradation-category rating

DCT	Discrete cosine transform
DEC	Dynamic equalization control
DECT	Digital enhanced cordless telecommunications
DFT	Discrete Fourier transform
DIVA	Digital interactive virtual acoustics
DRAM	Dynamic random access memory
DSP	Digital signal processing or digital signal processor
DTS	Digital Theater Systems Inc.
DVC	Dynamic volume control
DWT	Discrete wavelet transform
EDC	Energy decay curve
EMDF	Extended multi-delay filter
ERB	Equivalent rectangular bandwidth
ERLE	Echo return loss enhancement
ETSI	European telecommunications standards institute
FB	Filter bank
FBE	Filter-bank equalizer
FBSM	Filter-bank summation method
FDAF	Frequency domain adaptive filter
FFT	Fast Fourier transform
FIR	Finite impulse response
GAL	Gradient adaptive lattice
GCC	Generalized cross correlation
GDCT	Generalized discrete cosine transform
GDFT	Generalized discrete Fourier transform
GMAF	Generalized multi-delay filter
GSC	Generalized sidelobe canceller
GSM	Global system for mobile communications
HERB	Harmonicity-based dereverberation
HINT	Hearing in noise test
HMM	Hidden Markov model
HOS	Higher-order statistics
HP	Highpass
HRIR	Head related impulse response
HRTF	Head related transfer function
HTK	Hidden Markov model toolkit
ICA	Independent component analysis
ICC	In-car communication
IDEC	Individual dynamic equalization control
IDFT	Inverse discrete Fourier transform
IDVC	Individual dynamic volume control
IEC	International electrotechnical commission
IFFT	Inverse fast Fourier transform
IHC	Inner hair cell
IID	Interaural intensity difference or independent identically distributed

IIR	Infinite impulse response
IMCRA	Improved minima controlled recursive averaging
INMD	In-service non-intrusive measurement device
IRS	Intermediate reference system
ISDN	Integrated-services digital network
ISO	International standardization organization
ITD	Interaural time difference
ITU	International telecommunication union
IWDFT	Inverse warped discrete Fourier transform
KEMAR	Knowles electronic manikin for acoustic research
LAR	Log-area ratio
LDF	Low delay filter
LEM	Loudspeaker enclosure microphone
LMS	Least mean square
LOT	Listening-only test
LP	Linear prediction or lowpass
LPC	Linear predictive coding
LPTV	Linear periodically time-variant
LQ	Listing quality
LS	Least squares
LSA	Log spectral amplitude
LSD	Log spectral distance
LTI	Linear time-invariant
MA	Moving average
MAP	Maximum a posteriori
MDF	Multi-delay filter
MDS	Mulit-dimensional scaling
MFCC	Mel filtered cepstral coefficient
MIMO	Multiple-input multiple-output
MINT	Multiple input/output inverse theorem
MOS	Mean-opinion score
MOST	Media oriented systems transport
MSC	Magnitude-squared coherence
MVDR	Minimum variance distortionless response
NLMS	Normalized least mean square
NPR	Near-perfect reconstruction
NR	Noise reduction
NS	Noise suppression
OEM	Original equipment manufacturer
OHC	Outer hair cell
OM-LSA	Optimally-modified log spectral amplitude
PARCOR	Partial correlation (coefficient)
PAMS	Perceptual analysis measurement system
PBFDAF	Partitioned block frequency domain adaptive filter
PC	Personal computer

PCM	Pulse code modulation
PDF	Probability density function
PESQ	Perceptual evaluation of speech quality
PHAT	Phase transform
POTS	Plain old telephone system
PPN	Polyphase network
PR	Perfect reconstruction
PSD	Power spectral density
PSQM	Perceptual speech quality measure
QMF	Quadrature mirror filter
RAM	Random access memory
RASTA	Relative spectra
REMOS	Reverberation modeling for speech recognition
RES	Residual echo suppression
RF	Radio frequency
RIR	Room impulse response
RLS	Recursive least-squares
RPM	Revolutions per minute
RS	Reverberation suppression
SAEC	Stereo acoustic echo cancellation
SBC	Subband coding
SC	Sylvester constraint
SD	Semantic differential or spectral distance
SF	Smoothing filter
SFM	Spectral flatness measure
SIMO	Single-input multiple-output
SIR	Signal-to-interference ratio
SIRP	Spherically invariant random process
SNR	Signal-to-noise ratio
SRA	Statistical room acoustics
SRR	Signal-to-reverberation ratio
SOS	Second-order statistics
SPIN	Speech perception in noise
SPL	Sound pressure level
SQET	Speech-quality evaluation tool
SRAM	Static random access memory
STFT	Short-time Fourier transform
TBQ	Total background quality
TCM	Target cancellation module
TDOA	Time difference of arrival
TFRM	Tolerance function of raster match
TFTM	Tolerance function of triple match
TIMIT	Texas Instruments (TI) and Massachusetts Institute of Technology (MIT)
TOSQA	Telecommunication objective speech quality assessment
TRINICON	Triple-N independent component analysis for convolutive mixtures

TSQ	Total signal quality
TWRM	Tolerance width of raster match
TWTM	Tolerance width of triple match
VAD	Voice activity detection
VDA	Verband der Automobilindustrie (German, stands for *German association of the automotive industry*)
VoIP	Voice over internet protocol
WDFT	Warped discrete Fourier transform

Part I

Introduction

Part II

Speech Enhancement

Part III

Echo Cancellation

Part IV

Signal and System Quality Evaluation

Part V

Multi-Channel Processing

Blind source
separation!convolutive
TRINICON

1

Convolutive Blind Source Separation for Noisy Mixtures

Robert Aichner[1], Herbert Buchner[2], and Walter Kellermann[3]

[1] Microsoft Corporation, Redmond, WA, USA †
[2] Deutsche Telekom Laboratories, Technical University Berlin, Germany †
[3] University of Erlangen-Nuremberg, Germany

Convolutive blind source separation (BSS) is a promising technique for separating acoustic mixtures acquired by multiple microphones in reverberant environments. In contrast to conventional beamforming methods no a-priori knowledge about the source positions or sensor arrangement is necessary resulting in a greater versatility of the algorithms. In this contribution we will first review a general BSS framework called TRINICON which allows a unified treatment of broadband and narrowband BSS algorithms. Efficient algorithms will be presented and their high performance will be confirmed by experimental results in reverberant rooms. Subsequently, the BSS model will be extended by incorporating background noise. Commonly encountered realistic noise types are examined and, based on the resulting model, pre-processing methods for noise-robust BSS adaptation are investigated. Additionally, an efficient post-processing technique following the BSS stage, will be presented, which aims at simultaneous suppression of background noise and residual cross-talk. Combining these pre- or post-processing approaches with the algorithms obtained by the TRINICON framework yield versatile BSS systems which can be applied in adverse environments as will be demonstrated by experimental results.

1.1 Introduction

Acoustic blind source separation can be applied to scenarios where there are a number of point sources whose signals are picked up by several microphones. As each microphone is located at a different position, each sensor acquires a slightly different mixture of the original source signals. The goal of blind source separation is to recover the separated source signals from this set of

† The research underlying this work was performed while the authors were with Multimedia Communications and Signal Processing, University of Erlangen-Nuremberg.

Multiple-input multiple-output (MIMO)

sensor signals. The term “blind” stresses the fact that the source signals and the mixing system are assumed to be unknown and no information about the source positions and sensor arrangement is necessary. The fundamental assumption for BSS methods is that the original source signals are mutually statistically independent. In reality this assumption holds for a variety of signals, such as multiple speakers. Therefore, the problem of BSS refers to finding a demixing system whose outputs are statistically independent.

In reverberant environments delayed and attenuated versions of the source signals $s_q(n)$ are picked up by the microphones. Assuming point sources, this can be modeled by a mixing system consisting of finite impulse response (FIR) filters of length M given as

$$x_p(n) = \sum_{q=1}^{Q} \sum_{\kappa=0}^{M-1} h_{qp,\kappa} s_q(n-\kappa) + n_p(n), \tag{1.1}$$

where $h_{qp,\kappa}$, $\kappa = 0, \ldots, M-1$ denote the coefficients of the FIR filter model from the q-th source to the p-th sensor. In addition to the source signals, a noise signal $n_p(n)$ may be picked up by each sensor which contains both, background noise and sensor noise. In blind source separation, we are interested in finding a corresponding demixing system whose output signals $y_q(n)$ are described by

$$y_q(n) = \sum_{p=1}^{P} \sum_{\kappa=0}^{L-1} w_{pq,\kappa} x_p(n-\kappa). \tag{1.2}$$

The parameter L denotes the FIR filter length of the demixing filters $w_{pq,\kappa}$. The convolutive mixing model together with the demixing system is depicted as a block diagram in Fig. 1.1. From this it is obvious that BSS can be classified as a blind multiple-input multiple-output (MIMO) technique. Throughout this chapter, we regard the standard BSS model where the number Q of *potentially simultaneously active source signals* $s_q(n)$ is equal to the number

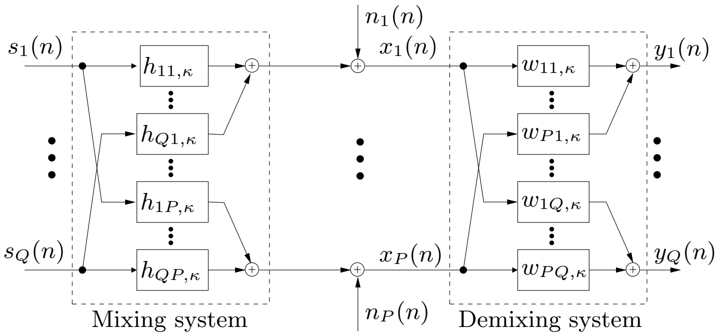


Fig. 1.1. Convolutive MIMO model for BSS.

of sensor signals $x_p(n)$, i.e., $Q = P$. It should be noted that in contrast to other BSS algorithms we do not assume prior knowledge about the exact number of active sources. Thus, even if the algorithms will be derived for the case $Q = P$, the number of simultaneously active sources may change throughout the application of the BSS algorithm and only the condition $Q \leq P$ has to be fulfilled. For $Q > P$ the demixing system cannot be computed directly so that usually the sparseness of the sources in transform domains, such as the discrete Fourier transform (DFT) domain, is exploited and subsequently time-frequency masking are applied to separate the sources. This research field is termed computational auditory scene analysis (CASA) and a recent overview on the state-of-the-art can be found, e.g., in [29, 84] or in Chap. **??** of this book. Alternative statistical approaches for the case of $P < Q$ are still in an early stage [87].

Nongaussianity
Higher-order statistics
Independent component analysis
Nonwhiteness
Nonstationarity

As pointed out above, the source signals $s_q(n)$ are assumed to be mutually statistically independent. For the case $P = Q$ considered in this chapter it was shown in [85] that merely utilizing second-order statistics (SOS) by decorrelating the output signals $y_q(n)$ does not lead to a separation of the sources. This implies that we have to force the output signals to become statistically decoupled up to joint moments of a certain order by using additional conditions. This can be realized by exploiting one of the following source signal properties:

(a) **Nongaussianity**. The probability density function (PDF) of an acoustic source signal $s_q(n)$ is in general not Gaussian. Thus, the nongaussianity can be exploited by using higher-order statistics (HOS) yielding a statistical decoupling of higher-order joint moments of the BSS output signals. BSS algorithms utilizing HOS are also termed independent component analysis (ICA) algorithms (e.g., [48, 77]).

(b) **Nonwhiteness**. Audio signals exhibit temporal dependencies which can be exploited by the BSS criterion. This means that the samples of *each* source signal are not independent along the time axis. However, the signal samples from different sources are *mutually* independent. Based on the assumption of mutual statistical independence for non-white sources, several algorithms can be found in the literature. There, mainly the nonwhiteness is exploited by simultaneous diagonalization of output correlation matrices over multiple time-lags, (e.g., [54, 67, 78, 81]).

(c) **Nonstationarity**. Audio signals are in general assumed to be nonstationary. Therefore, in most acoustic BSS applications nonstationarity of the source signals is exploited by simultaneous diagonalization of short-time output correlation matrices at different time instants (e.g., [50, 64, 73, 85]). The signals within each block, as necessary for estimating the correlation matrices, are usually assumed to be wide-sense stationary.

A simultaneous exploitation of two or even all three signal properties leads to improved results as was shown within the TRINICON framework [14, 16] which will be reviewed in the next section.

Permutation ambiguity
Filtering ambiguity
Deconvolution
Dereverberation

It should be pointed out that, as long as the concept of BSS is solely based on the assumption of mutual independence of the source signals, some ambiguities are unavoidable:

- Permutation ambiguity: The ordering of the separated sources cannot be determined.
- Filtering ambiguity: The estimated separated source signals can only be determined up to an arbitrary filtering operation.

The permutation ambiguity cannot be resolved without additional a-priori information. However, if, e.g., the sensor positions are known, then the position of each separated source can be determined from the demixing system [18]. For some applications this may be sufficient for solving the permutation problem.

The filtering ambiguity is caused by the fact that in general, BSS approaches do not aim at blind dereverberation which would lead to a deconvolution of the mixing system, i.e., at a recovery of the original source signals up to an arbitrary scaling factor and a constant delay. Blind dereverberation is a more challenging task as it requires to distinguish between the temporal correlations introduced by the vocal tract of the human speaker and the correlations originating from the reverberation of the room. This was addressed in the extension of the TRINICON framework to blind dereverberation in [15]. However, even if we do not strive for solving the dereverberation problem in BSS it is still desirable to avoid the arbitrariness of the filtering operation in blind source separation. Fortunately, it can be shown [8, 20] that the filtering ambiguity reduces to a scaling ambiguity, if the demixing filter length L is chosen less or equal to the optimum BSS demixing filter length $L_{\mathrm{opt}} = \frac{(Q-1)(M-1)+1}{P-Q+1}$. Another popular approach to avoid the arbitrary filtering is to apply a constraint which minimizes the distortion introduced by the demixing system of the BSS algorithm. Thereby, the q-th separated source $y_q(n)$ is constrained to be equal to the component of the desired source $s_q(n)$ picked up, e.g., at the q-th microphone. This is done by back-projecting the estimated sources to the sensors or by introducing a constrained optimization scheme [49, 65]. In the following we disregard these ambiguities and first concentrate on the fundamental BSS problem for convolutive acoustic mixtures and then extend our treatment to noisy mixtures.

The rest of the chapter is structured as follows: In the next section the TRINICON framework which is based on a generic time-domain optimization criterion accounting for all three signal properties is reviewed. The minimization of the criterion leads to a natural gradient algorithm which exhibits a so-called Sylvester constraint. Subsequently, several approximations are discussed yielding various efficient BSS algorithms and experimental results in reverberant environments are given. In Sec. 1.3 the framework is extended to noisy environments. First, a model for background noise is discussed. Based on this model several pre-processing methods and a post-processing approach are presented which complement the BSS algorithms derived from the TRINICON framework. Especially the most promising post-processing scheme is discussed in detail and experimental results demonstrate the increased versatility of the complemented BSS algorithms.

Sylvester!structure

1.2 Blind Source Separation for Acoustic Mixtures Based on the TRINICON Framework

In this section, we introduce, based on a compact matrix notation, a generic convolutive BSS framework which allows the simultaneous exploitation of the three signal properties. Several efficient algorithms are presented which can be derived from the optimization criterion of the framework and which allow real-time separation of multiple sources in reverberant environments. Moreover, links to well-known algorithms in the literature are illustrated.

1.2.1 Matrix Formulation

From the convolutive MIMO model illustrated in Fig. 1.1 it can be seen that the output signals $y_q(n)$ are obtained by convolving the input signals $x_p(n)$ with the demixing filter coefficients $w_{pq,\kappa}$, $\kappa = 0, \ldots, L-1$. For an algorithm which utilizes the nonwhiteness property of the source signals accounting for $D-1$ time-lags, a memory containing the current and the previous $D-1$ output signal values $y_q(n), \ldots, y_q(n-D+1)$ has to be introduced. The linear convolution yielding the D output signal values can be formulated using matrix-vector notation as

$$\boldsymbol{y}_q(n) = \sum_{p=1}^{P} \boldsymbol{W}_{pq}^{\mathrm{T}} \boldsymbol{x}_p(n), \tag{1.3}$$

with the column vectors $\boldsymbol{x}_p$ and $\boldsymbol{y}_q$ given as[1]

$$\boldsymbol{x}_p(n) = \left[x_p(n), \ldots, x_p(n-2L+1)\right]^{\mathrm{T}}, \tag{1.4}$$

$$\boldsymbol{y}_q(n) = \left[y_q(n), \ldots, y_q(n-D+1)\right]^{\mathrm{T}}. \tag{1.5}$$

To express the linear convolution as a matrix-vector product, the $2L \times D$ matrix $\boldsymbol{W}_{pq}$ exhibits a Sylvester structure that contains all L coefficients of the respective demixing filter in each column:

[1] With respect to efficient DFT-domain implementations the vector $\boldsymbol{x}_p$ contains $2L$ sensor signal samples instead of the $L+D-1$ samples required for the linear convolution $(1 \leq D \leq L)$.

Mutual information
TRINICON

$$\boldsymbol{W}_{pq} = \begin{bmatrix} w_{pq,0} & 0 & \cdots & 0 \\ w_{pq,1} & w_{pq,0} & \ddots & \vdots \\ \vdots & w_{pq,1} & \ddots & 0 \\ w_{pq,L-1} & \vdots & \ddots & w_{pq,0} \\ 0 & w_{pq,L-1} & \ddots & w_{pq,1} \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & w_{pq,L-1} \\ 0 & \cdots & 0 & 0 \\ \vdots & \cdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 \end{bmatrix} . \tag{1.6}$$

It can be seen that for the general case, $1 \leq D \leq L$, the last $L - D + 1$ rows of $\boldsymbol{W}_{pq}$ are padded with zeros to ensure compatibility with the length of $\boldsymbol{x}_p(n)$ which was chosen to $2L$. Note that for $D = 1$, Eq. 1.3 simplifies to the well-known vector formulation of a convolution, as it is used extensively in the literature on supervised adaptive filtering, e.g., [41]. Finally, to allow a convenient notation we combine all channels and thus, we can write Eq. 1.3 compactly as

$$\boldsymbol{y}(n) = \boldsymbol{W}^{\mathrm{T}} \boldsymbol{x}(n), \tag{1.7}$$

with

$$\boldsymbol{x}(n) = \left[\boldsymbol{x}_1^{\mathrm{T}}(n), \ldots, \boldsymbol{x}_P^{\mathrm{T}}(n)\right]^{\mathrm{T}}, \tag{1.8}$$

$$\boldsymbol{y}(n) = \left[\boldsymbol{y}_1^{\mathrm{T}}(n), \ldots, \boldsymbol{y}_P^{\mathrm{T}}(n)\right]^{\mathrm{T}}, \tag{1.9}$$

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{W}_{11} & \cdots & \boldsymbol{W}_{1P} \\ \vdots & \ddots & \vdots \\ \boldsymbol{W}_{P1} & \cdots & \boldsymbol{W}_{PP} \end{bmatrix}, \tag{1.10}$$

with $\boldsymbol{W}$ exhibiting a blockwise Sylvester structure.

1.2.2 Optimization Criterion and Coefficient Update

As pointed out before, we aim at an optimization criterion simultaneously exploiting the three signal properties nonstationarity, nonwhiteness, and nongaussianity. Therefore, based on a generalization of Shannon's mutual information [27], the following optimization criterion was defined in [14] and was termed " **TRI**ple-**N**-**I**ndependent component analysis for **CON**volutive mixtures" (**TRINICON**) as it simultaneously accounts for the three fundamental properties **N**onwhiteness, **N**onstationarity, and **N**ongaussianity:

$$\mathcal{J}(m,\boldsymbol{W}) = \sum_{i=0}^{\infty} \beta(i,m) \frac{1}{N} \sum_{j=0}^{N-1} \left\{ \log \frac{\hat{p}_{y,PD}\big(\boldsymbol{y}(iL+j)\big)}{\prod_{q=1}^{P} \hat{p}_{y_q,D}\big(\boldsymbol{y}_q(iL+j)\big)} \right\} . \quad (1.11)$$

Nongaussianity
Nonwhiteness
Multivariate probability density function
Nonstationarity

The variable $\hat{p}_{y_q,D}(\cdot)$ is the estimated or assumed multivariate probability density function (PDF) for channel q of dimension D and $\hat{p}_{y,PD}(\cdot)$ is the joint PDF of dimension PD over all channels. The usage of PDFs allows to exploit the *nongaussianity* of the signals. Furthermore, the multivariate structure of the PDFs, which is given by the memory length D, i.e., the number of time-lags, models the *nonwhiteness* of the P signals with D chosen to $1 \leq D \leq L$. The expectation operator of the mutual information [27] is replaced in Eq. 1.11 by a short-time estimate of the multivariate PDFs using N time instants. To allow for a proper estimation of the multivariate PDFs the averaging has to be done in general for $N > PD$ time instants. The block indices i, m refer to the blocks which are underlying to the statistical estimation of the multivariate PDFs. For each output signal block $\boldsymbol{y}_q(iL+j)$ containing D samples a sensor signal block of length $2L$ is required according to Eq. 1.4. The *nonstationarity* is taken into account by a weighting function $\beta(i,m)$ with the block indices i, m and with finite support. The weighting function is normalized according to

$$\sum_{i=0}^{\infty} \beta(i,m) = 1 \,, \quad (1.12)$$

and allows offline, online, and block-online implementations of the algorithms [16]. As an example,

$$\beta(i,m) = \begin{cases} (1-\lambda)\lambda^{m-i}, & \text{for } 0 \leq i \leq m, \\ 0, & \text{else,} \end{cases} \quad (1.13)$$

leads to an efficient online version allowing for tracking in time-variant environments. The forgetting factor λ is usually chosen close to, but less than 1. A robust block-online adaptation was discussed in detail in [6].

The approach followed here is carried out with overlapping data blocks as the sensor signal blocks of length $2L$ are shifted only by L samples due to the time index iL in Eq. 1.11. Analogously to supervised block-based adaptive filtering [41], this increases the convergence rate and reduces the signal delay. If further overlapping is desired, then the time index iL in Eq. 1.11 is simply replaced by iL/α. The overlap factor α with $1 \leq \alpha \leq L$ should be chosen suitably to obtain integer values for the time index.

The derivation of the gradient with respect to the demixing filter weights $w_{pq,\kappa}$ for $p, q = 1, \ldots, P$ and $\kappa = 0, \ldots, L-1$ can be expressed compactly in matrix notation by defining the matrix $\check{\boldsymbol{W}}$ given as

Sylvester!matrix
Sylvester constraint

$$\check{\boldsymbol{W}} = \begin{bmatrix} \boldsymbol{w}_{11} & \cdots & \boldsymbol{w}_{1P} \\ \vdots & \ddots & \vdots \\ \boldsymbol{w}_{P1} & \cdots & \boldsymbol{w}_{PP} \end{bmatrix},$$

which is composed of the column vectors $\boldsymbol{w}_{pq}$ containing the demixing filter coefficients

$$\boldsymbol{w}_{pq} = \left[w_{pq,0}, \ldots, w_{pq,L-1}\right]^{\mathrm{T}}. \tag{1.14}$$

Then the gradient with respect to the P^2L demixing filter coefficients can be expressed compactly as

$$\nabla_{\check{\boldsymbol{W}}} \mathcal{J}(m, \boldsymbol{W}) = \frac{\partial \mathcal{J}(m, \boldsymbol{W})}{\partial \check{\boldsymbol{W}}}. \tag{1.15}$$

With an iterative optimization procedure, the current demixing matrix is obtained by the recursive update equation

$$\check{\boldsymbol{W}}(m) = \check{\boldsymbol{W}}(m-1) - \mu \Delta \check{\boldsymbol{W}}(m), \tag{1.16}$$

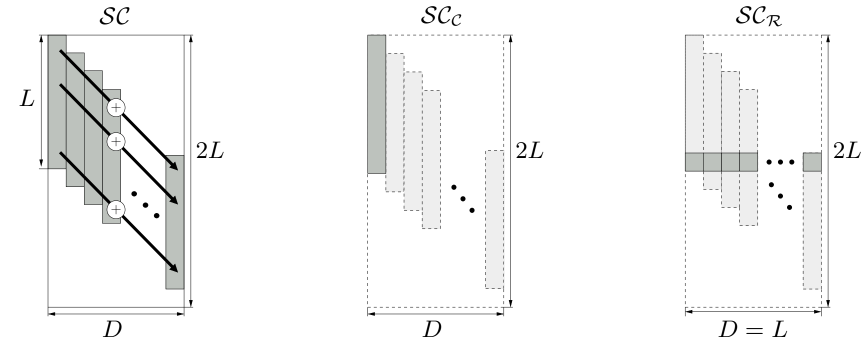
where μ is a stepsize parameter, and $\Delta\check{\boldsymbol{W}}(m)$ is the update which is set equal to $\nabla_{\check{\boldsymbol{W}}} \mathcal{J}(m, \boldsymbol{W})$ for gradient descent adaptation.

In order to calculate the gradient (Eq. 1.15), the TRINICON optimization criterion $\mathcal{J}(m, \boldsymbol{W})$ given in Eq. 1.11 has to be expressed in terms of the demixing filter coefficients $w_{pq,\kappa}$. This can be done by inserting the definition of the linear convolution $\boldsymbol{y} = \boldsymbol{W}^{\mathrm{T}}\boldsymbol{x}$ given in Eq. 1.7 into $\mathcal{J}(m, \boldsymbol{W})$ and subsequently transforming the output signal PDF $\hat{p}_{y,PD}(\boldsymbol{y}(iL+j))$ into the PD-dimensional input signal PDF $\hat{p}_{x,PD}(\cdot)$ using the Sylvester matrix $\boldsymbol{W}$, which is considered as a mapping matrix for this linear transformation [71]. This leads to an expression of the optimization criterion 1.11 with respect to the Sylvester matrix $\boldsymbol{W}$. To be able to take the derivative with respect to $\check{\boldsymbol{W}}$ instead of the Sylvester matrix $\boldsymbol{W}$, the chain rule for the derivative of a scalar function with respect to a matrix [40] was applied to the gradient (Eq. 1.15) in [20]. There, it was shown that the chain rule leads to a $\mathcal{S}$ylvester $\mathcal{C}$onstraint operator ($\mathcal{SC}$) which relates the gradient with respect to $\check{\boldsymbol{W}}$ and with respect to $\boldsymbol{W}$ as

$$\nabla_{\check{\boldsymbol{W}}} \mathcal{J}(m, \boldsymbol{W}) = \mathcal{SC}\left\{\nabla_{\boldsymbol{W}} \mathcal{J}(m, \boldsymbol{W})\right\}. \tag{1.17}$$

The Sylvester constraint operator $\mathcal{SC}$ is illustrated for the pq-th submatrix of $\nabla_{\boldsymbol{W}} \mathcal{J}(m, \boldsymbol{W})$ in the left plot of Fig. 1.2 where it can be seen that it corresponds (up to a scaling by the constant factor D) to an arithmetic average over the elements on each diagonal of the $2L \times D$ submatrices of the gradient $\nabla_{\boldsymbol{W}} \mathcal{J}(m, \boldsymbol{W})$. Thus, the $2PL \times PD$ gradient $\nabla_{\boldsymbol{W}} \mathcal{J}(m, \boldsymbol{W})$ will be reduced to the $PL \times P$ gradient $\nabla_{\check{\boldsymbol{W}}} \mathcal{J}(m, \boldsymbol{W})$.

To reduce computational complexity, two efficient approximated versions of the Sylvester constraint $\mathcal{SC}$ (see Fig. 1.2) were discussed in [6] leading to two different classes of algorithms:



Natural gradient

Fig. 1.2. The Sylvester constraint ($\mathcal{SC}$) and two popular approximations denoted as the column Sylvester constraint $\mathcal{SC}_\mathcal{C}$ and row Sylvester constraint $\mathcal{SC}_\mathcal{R}$ all illustrated for the gradient $\nabla_{\boldsymbol{W}_{pq}}\mathcal{J}(m, \boldsymbol{W})$ with respect to the pq-th submatrix $\boldsymbol{W}_{pq}$.

(1) Computing only the *first column* of each channel of the update matrix to obtain the new coefficient matrix $\check{\boldsymbol{W}}$. This method is denoted as $\mathcal{SC}_\mathcal{C}$.

(2) Computing only the *L-th row* of each channel of the update matrix to obtain the new coefficient matrix $\check{\boldsymbol{W}}$. This method is denoted as $\mathcal{SC}_\mathcal{R}$.

It can be shown that in both cases the update process is considerably simplified [6]. However, in general, both choices require some tradeoff in the algorithm performance. While simulations showed [4] that $\mathcal{SC}_\mathcal{C}$ may provide a potentially more robust convergence behaviour, it will not work for arbitrary source positions (e.g., in the case of two sources, they are required to be located in different half-planes with respect to the orientation of the microphone array), or for $P > 2$, which is in contrast to the more versatile $\mathcal{SC}_\mathcal{R}$ [4, 6]. Note that the choice of $\mathcal{SC}$ also determines the appropriate coefficient initialization [4, 6].

It is known that stochastic gradient descent, i.e., $\Delta\check{\boldsymbol{W}}(m) = \nabla_{\check{\boldsymbol{W}}}\mathcal{J}(m, \boldsymbol{W})$ suffers from slow convergence in many practical problems. In the BSS application the gradient and thus, the separation performance depends on the MIMO mixing system. Fortunately, a modification of the ordinary gradient, termed the *natural gradient* by Amari [9] and the *relative gradient* by Cardoso [21] (which is equivalent to the natural gradient in the BSS application) has been developed that largely removes all effects of an ill-conditioned mixing matrix, assuming an appropriate initialization of $\boldsymbol{W}$ and thus leads to better performance compared to the stochastic gradient descent. The idea of the relative gradient is based on the equivariance property. Generally speaking, an estimator behaves equivariantly if it produces estimates that, under data transformation, are transformed in the same way as the data [21]. In the context of BSS the key property of equivariant estimators is that they exhibit uniform performance, e.g., in terms of bias and variance, independently of the mixing system. In [17] the natural/relative gradient has been extended to the case of Sylvester matrices $\boldsymbol{W}$ which together with the Sylvester constraint

Multivariate score function

yields

$$\nabla^{\mathrm{NG}}_{\check{\boldsymbol{W}}} \mathcal{J}(m, \boldsymbol{W}) = \mathcal{SC}\left\{\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}} \nabla_{\boldsymbol{W}} \mathcal{J}(m, \boldsymbol{W})\right\}. \tag{1.18}$$

This leads to the following expression for the *HOS natural gradient*

$$\nabla^{\mathrm{NG}}_{\check{\boldsymbol{W}}} \mathcal{J}(m, \boldsymbol{W}) \\ = \mathcal{SC}\left\{\sum_{i=0}^{\infty} \beta(i,m)\boldsymbol{W}(i)\frac{1}{N}\sum_{j=0}^{N-1}\left\{\boldsymbol{y}(iL+j)\boldsymbol{\Phi}^{\mathrm{T}}\big(\boldsymbol{y}(iL+j)\big) - \boldsymbol{I}\right\}\right\}, \tag{1.19}$$

with the general weighting function $\beta(i,m)$ and the *multivariate score function* $\boldsymbol{\Phi}(\boldsymbol{y}(.))$ consisting of the stacked channel-wise multivariate score functions $\boldsymbol{\Phi}_q(\boldsymbol{y}_q(.))$, $q = 1, \ldots, P$ defined as

$$\boldsymbol{\Phi}\big(\boldsymbol{y}(iL+j)\big) = \left[\left(-\frac{\frac{\partial \hat{p}_{y_1,D}(\boldsymbol{y}_1(iL+j))}{\partial \boldsymbol{y}_1(iL+j)}}{\hat{p}_{y_1,D}\big(\boldsymbol{y}_1(iL+j)\big)}\right)^{\mathrm{T}}, \ldots, \left(-\frac{\frac{\partial \hat{p}_{y_P,D}(\boldsymbol{y}_P(iL+j))}{\partial \boldsymbol{y}_P(iL+j)}}{\hat{p}_{y_P,D}\big(\boldsymbol{y}_P(iL+j)\big)}\right)^{\mathrm{T}}\right]^{\mathrm{T}} \\ := \left[\boldsymbol{\Phi}_1^{\mathrm{T}}\big(\boldsymbol{y}_1(iL+j)\big), \ldots, \boldsymbol{\Phi}_P^{\mathrm{T}}\big(\boldsymbol{y}_P(iL+j)\big)\right]^{\mathrm{T}}. \tag{1.20}$$

The update in Eq. 1.19 represents a so-called holonomic algorithm as it imposes the constraint $\boldsymbol{y}(iL+j)\boldsymbol{\Phi}^{\mathrm{T}}(\boldsymbol{y}(iL+j)) = \boldsymbol{I}$ on the magnitudes of the recovered signals. However, when the source signals are nonstationary, these constraints may force a rapid change in the magnitude of the demixing matrix which in turn leads to numerical instabilities in some cases (see, e.g., [25]). By replacing $\boldsymbol{I}$ in Eq. 1.19 with the term $\mathrm{bdiag}\{\boldsymbol{y}(iL+j)\boldsymbol{\Phi}^{\mathrm{T}}(\boldsymbol{y}(iL+j))\}$ the constraint on the magnitude of the recovered signals can be avoided. This is termed the *nonholonomic natural gradient* algorithm which is given as

$$\nabla^{\mathrm{NG}}_{\check{\boldsymbol{W}}} \mathcal{J}(m, \boldsymbol{W}) = \mathcal{SC}\left\{\sum_{i=0}^{\infty} \beta(i,m)\boldsymbol{W}(i)\frac{1}{N}\sum_{j=0}^{N-1}\left\{\boldsymbol{y}(iL+j)\boldsymbol{\Phi}^{\mathrm{T}}\big(\boldsymbol{y}(iL+j)\big) \right.\right. \\ \left.\left. - \,\mathrm{bdiag}\left\{\boldsymbol{y}(iL+j)\boldsymbol{\Phi}^{\mathrm{T}}\big(\boldsymbol{y}(iL+j)\big)\right\}\right\}\right\}. \tag{1.21}$$

Here, the bdiag operator sets all cross-channel terms to zero. Due to the improved convergence behaviour and the nonstationary nature of acoustic signals the remainder of this chapter will focus on the nonholonomic algorithm (Eq. 1.21) based on the natural gradient.

1.2.3 Approximations Leading to Special Cases

The natural gradient update (Eq. 1.21) rule provides a very general basis for BSS of convolutive mixtures. However, to apply it to real-world scenarios, the multivariate score function (Eq. 1.20) has to be estimated, i.e., we have to estimate P multivariate PDFs $\hat{p}_{y_q,D}(\boldsymbol{y}_q(iL+j))$, $q=1,\ldots,P$ of dimension D. In general, this is a very challenging task, as it effectively requires estimation of all possible higher-order cumulants for a set of D output samples, where D may be on the order of several hundred or thousand in real acoustic environments.

As first shown in [14] we will present in Sec. 1.2.3.1 an efficient solution for the problem of estimating the multivariate score function by assuming so-called *spherically invariant random processes* (SIRPs). Moreover, efficient realizations based on second-order statistics will be derived in Sec. 1.2.3.2 by utilization of the multivariate Gaussian PDF.

Multivariate probability density function
Multivariate score function
Spherically invariant random process (SIRP)
Higher-order statistics
Univariate probability density function
Multivariate Gaussian probability density function

1.2.3.1 Higher-Order Statistics Realization Based on Multivariate PDFs

Early experimental measurements [28] indicated that the PDF of speech signals in the time domain can be approximated by exponential distributions such as the Gamma or Laplacian PDF. Later on, a special class of multivariate PDFs based on the assumption of SIRPs was used in [13] to model bandlimited telephone speech. The SIRP model is representative for a wide class of stochastic processes [35,74,89] and is very attractive since multivariate PDFs can be derived analytically from the corresponding univariate probability density function together with the correlation matrices covering multiple time-lags. The correlation matrices can be estimated from the data while for the univariate PDF appropriate models can be assumed or the univariate PDF can be estimated based on parameterized representations, such as the Gram-Charlier or Edgeworth expansions [48].

The general model of a zero-mean non-white SIRP of D-th order for channel q is given by [13]

$$\begin{aligned}&\hat{p}_{y_q,D}\big(\boldsymbol{y}_q(iL+j)\big)\\&\quad=\frac{1}{\sqrt{\pi^D\det(\boldsymbol{R}_{\boldsymbol{y}_q\boldsymbol{y}_q}(i))}}f_{y_q,D}\left(\boldsymbol{y}_q^{\mathrm{T}}(iL+j)\boldsymbol{R}_{\boldsymbol{y}_q\boldsymbol{y}_q}^{-1}(i)\boldsymbol{y}_q(iL+j)\right)\end{aligned}\tag{1.22}$$

with the $D\times D$ correlation matrix given as

$$\boldsymbol{R}_{\boldsymbol{y}_p\boldsymbol{y}_q}(i)=\frac{1}{N}\sum_{j=0}^{N-1}\boldsymbol{y}_p(iL+j)\,\boldsymbol{y}_q^{\mathrm{T}}(iL+j),\tag{1.23}$$

and the function $f_{y_q,D}(\cdot)$ depending on the chosen univariate PDF. As the best known example, the multivariate Gaussian can be viewed as a special case of the class of SIRPs. The multivariate PDFs are completely characterized by the scalar function $f_{y_q,D}(\cdot)$ and $\boldsymbol{R}_{\boldsymbol{y}_q\boldsymbol{y}_q}$. Due to the quadratic form $\boldsymbol{y}_q^{\mathrm{T}}\boldsymbol{R}_{\boldsymbol{y}_q\boldsymbol{y}_q}^{-1}\boldsymbol{y}_q$, the PDF is spherically invariant which means for the bivariate case ($D=2$)

Multivariate score function

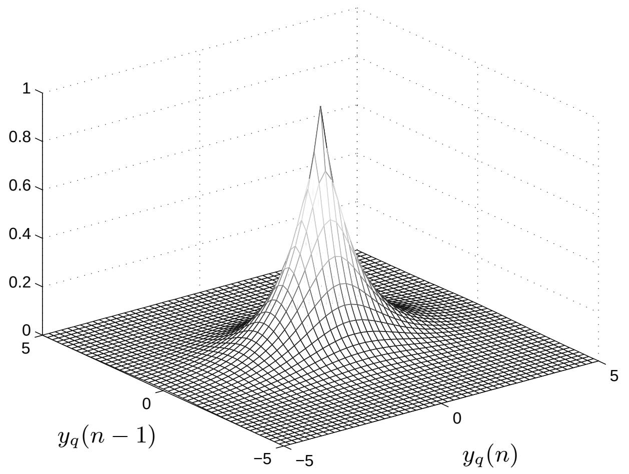


Fig. 1.3. Illustration of a bivariate SIRP PDF (i.e., $D = 2$).

that independent of the choice of $f_{y_q,D}(\cdot)$ the bivariate PDFs based on the SIRP model exhibit ellipsoidal or circular contour lines (see Fig. 1.3). The function $f_{y_q,D}(\cdot)$ is determined by the choice of the univariate PDF and can be calculated by using the so-called Meijer's G-functions as detailed in [13].

By introducing SIRPs into the BSS optimization criterion we obtain a considerably simplified expression for the multivariate score function (Eq. 1.20) as first presented in [14]. After applying the chain rule to Eq. 1.22, the multivariate score function for the q-th channel can be expressed as

$$\begin{aligned}\boldsymbol{\Phi}_q\big(\boldsymbol{y}_q(iL+j)\big) &= -\frac{\frac{\partial \hat{p}_{y_q,D}(\boldsymbol{y}_q(iL+j))}{\partial \boldsymbol{y}_q(iL+j)}}{\hat{p}_{y_q,D}(\boldsymbol{y}_q(iL+j))} \\ &= 2\underbrace{\left[-\frac{\frac{\partial f_{y_q,D}(u_q(iL+j))}{\partial u_q(iL+j)}}{f_{y_q,D}(u_q(iL+j))}\right]}_{:=\phi_{y_q,D}(u_q(iL+j))} \boldsymbol{R}^{-1}_{\boldsymbol{y}_q\boldsymbol{y}_q}(i)\,\boldsymbol{y}_q(iL+j). \quad (1.24)\end{aligned}$$

For convenience, we call the scalar function $\phi_{y_q,D}(u_q(iL+j))$ the *SIRP score* of channel q and the scalar argument given as the quadratic form is defined as

$$u_q(iL+j) = \boldsymbol{y}_q^{\mathrm{T}}(iL+j)\,\boldsymbol{R}^{-1}_{\boldsymbol{y}_q\boldsymbol{y}_q}(i)\,\boldsymbol{y}_q(iL+j). \quad (1.25)$$

From Eq. 1.24 it can be seen that the estimation of multivariate PDFs reduces to an estimation of the correlation matrix together with a computation of the

Spherically symmetric multivariate Laplacian PDF

Natural gradient

SIRP score which can be determined by choosing suitable models for the multivariate SIRP PDF.

In [13, 34] it was shown that the *spherically symmetric multivariate Laplacian PDF* which exhibits Laplacian marginals is a good model for long-term properties of speech signals in the time-domain. A derivation of the multivariate Laplacian based on SIRPs can be found in, e.g., [13, 31, 53] and leads to Eq. 1.22 with the function $f_{y_q,D}(u_q(iL+j))$ given as

$$f_{y_q,D}\big(u_q(iL+j)\big) = \left(\frac{1}{\sqrt{2u_q(iL+j)}}\right)^{D/2-1} K_{D/2-1}\left(\sqrt{2u_q(iL+j)}\right). \tag{1.26}$$

where $K_\nu(\cdot)$ denotes the ν-th order modified Bessel function of the second kind. The SIRP score for the multivariate Laplacian SIRP PDF can be straightforwardly derived by using the relation for the derivative of a ν-th order modified Bessel function of the second kind given as [1]

$$\frac{\partial K_\nu\left(\sqrt{2u_q}\right)}{\partial\sqrt{2u_q}} = \frac{\nu}{\sqrt{2u_q}} K_\nu\left(\sqrt{2u_q}\right) - K_{\nu+1}\left(\sqrt{2u_q}\right), \tag{1.27}$$

and is obtained as

$$\phi_{y_q,D}\big(u_q(iL+j)\big) = \frac{1}{\sqrt{2u_q(iL+j)}} \frac{K_{D/2}\left(\sqrt{2u_q(iL+j)}\right)}{K_{D/2-1}\left(\sqrt{2u_q(iL+j)}\right)}. \tag{1.28}$$

It should be noted that the formulation of Eq. 1.28 in [14] is slightly different but equivalent. In practical implementations the ν-th order modified Bessel function of the second kind $K_\nu(\sqrt{2u_q})$ may be approximated by [1]

$$K_\nu\left(\sqrt{2u_q}\right) = \sqrt{\frac{\pi}{2\sqrt{2u_q}}}\, e^{-\sqrt{2u_q}} \left(1 + \frac{4\nu^2-1}{8\sqrt{2u_q}} + \frac{(4\nu^2-1)(4\nu^2-9)}{2!(8\sqrt{2u_q})^2} + \ldots\right). \tag{1.29}$$

Having derived the multivariate score function (Eq. 1.24) for the SIRP model, we can now insert it into the generic HOS natural gradient update equation with its nonholonomic extension (Eq. 1.21) and will find several attractive properties that lead to significant reductions in computational complexity relative to the general case. Considering the fact that the auto-correlation matrices are symmetric so that $(\boldsymbol{R}^{-1}_{\boldsymbol{y}_q\boldsymbol{y}_q})^{\mathrm{T}} = \boldsymbol{R}^{-1}_{\boldsymbol{y}_q\boldsymbol{y}_q}$ leads to the following expression for the *nonholonomic HOS-SIRP natural gradient*:

$$\begin{aligned}&\nabla^{\mathrm{NG}}_{\check{\boldsymbol{W}}}\mathcal{J}(m,\boldsymbol{W})\\ &= \mathcal{SC}\left\{2\sum_{i=0}^{\infty}\beta(i,m)\boldsymbol{W}(i)\Big[\boldsymbol{R}_{\boldsymbol{y}\phi(\boldsymbol{y})}(i) - \mathrm{bdiag}\Big\{\boldsymbol{R}_{\boldsymbol{y}\phi(\boldsymbol{y})}(i)\Big\}\Big]\mathrm{bdiag}^{-1}\Big\{\boldsymbol{R}_{\boldsymbol{y}\boldsymbol{y}}(i)\Big\}\right\}\end{aligned} \tag{1.30}$$

Second-order statistics
Multivariate Gaussian PDF
Multivariate score function

with the second-order correlation matrix $\boldsymbol{R}_{\boldsymbol{yy}}$ consisting of the channel-wise submatrices $\boldsymbol{R}_{\boldsymbol{y}_p\boldsymbol{y}_q}$ defined in Eq. 1.23 and $\boldsymbol{R}_{\boldsymbol{y}\phi(\boldsymbol{y})}$ consisting of the channel-wise submatrices $\boldsymbol{R}_{\boldsymbol{y}_p\phi(\boldsymbol{y}_q)}$ given as

$$\boldsymbol{R}_{\boldsymbol{y}_p\phi(\boldsymbol{y}_q)}(i) = \frac{1}{N}\sum_{j=0}^{N-1} \boldsymbol{y}_p(iL+j)\phi_{y_q,D}\big(u_q(iL+j)\big)\boldsymbol{y}_q^{\mathrm{T}}(iL+j). \qquad (1.31)$$

The SIRP score $\phi_{y_q,D}(\cdot)$ of channel q which is a scalar value function causes a weighting of the correlation matrix in Eq. 1.31. In Eq. 1.30 only channel-wise submatrices have to be inverted so that it is sufficient to choose $N > D$ instead of $N > PD$ for the estimation of $\boldsymbol{R}_{\boldsymbol{yy}}(i)$ and $\boldsymbol{R}_{\boldsymbol{y}\phi(\boldsymbol{y})}$. Moreover, from the update equation 1.30, it can be seen that the SIRP model leads to an inherent normalization by the auto-correlation submatrices. This becomes especially obvious if the update (Eq. 1.30) is written explicitly for a 2-by-2 MIMO system leading to

$$\begin{aligned}&\nabla^{\mathrm{NG}}_{\check{\boldsymbol{W}}}\mathcal{J}(m,\boldsymbol{W})\\ &= \mathcal{SC}\left\{2\sum_{i=0}^{\infty}\beta(i,m)\boldsymbol{W}(i)\begin{bmatrix} \boldsymbol{0} & \boldsymbol{R}_{\boldsymbol{y}_1\phi(\boldsymbol{y}_2)}(i)\boldsymbol{R}^{-1}_{\boldsymbol{y}_2\boldsymbol{y}_2}(i) \\ \boldsymbol{R}_{\boldsymbol{y}_2\phi(\boldsymbol{y}_1)}(i)\boldsymbol{R}^{-1}_{\boldsymbol{y}_1\boldsymbol{y}_1}(i) & \boldsymbol{0}\end{bmatrix}\right\}.\end{aligned} \qquad (1.32)$$

The normalization is important as it provides good convergence even for correlated signals such as speech and also for a large number of filter taps. The normalization is similar as in the recursive least-squares (RLS) algorithm in supervised adaptive filtering where also the inverse of the auto-correlation matrix is computed [41]. To obtain efficient implementations, the normalization by the computationally demanding inverse of the $D \times D$ matrix can be approximated in several ways as shown in Sec. 1.2.4 and outlined in Sec. 1.2.5.

1.2.3.2 Second-Order Statistics Realization Based on the Multivariate Gaussian PDF

Using the model of the multivariate Gaussian PDF leads to a second-order realization of the BSS algorithm utilizing the nonstationarity and the nonwhiteness of the source signals. The multivariate Gaussian PDF

$$\hat{p}_{y_q,D}\big(\boldsymbol{y}_q(iL+j)\big) = \frac{1}{\sqrt{(2\pi)^D \det\big(\boldsymbol{R}_{\boldsymbol{y}_q\boldsymbol{y}_q}(i)\big)}}\, e^{-\frac{1}{2}\boldsymbol{y}_q^{\mathrm{T}}(iL+j)\boldsymbol{R}^{-1}_{\boldsymbol{y}_q\boldsymbol{y}_q}(i)\boldsymbol{y}_q(iL+j)} \qquad (1.33)$$

is inserted in the expression for the multivariate score function (Eq. 1.20) whose elements reduce to

$$\boldsymbol{\Phi}_q\big(\boldsymbol{y}_q(iL+j)\big) = \boldsymbol{R}^{-1}_{\boldsymbol{y}_q\boldsymbol{y}_q}(i)\,\boldsymbol{y}_q(iL+j)\,. \qquad (1.34)$$

Natural gradient
Nonwhiteness
Nonstationarity

Inserting Eq. 1.34 into the natural gradient update (Eq. 1.19) yields the *SOS natural gradient*:

$$\begin{aligned}&\nabla_{\check{\boldsymbol{W}}}^{\mathrm{NG}}\mathcal{J}(m,\boldsymbol{W})\\&=\mathcal{SC}\left\{\sum_{i=0}^{\infty}\beta(i,m)\boldsymbol{W}(i)\Big[\boldsymbol{R}_{\boldsymbol{yy}}(i)-\mathrm{bdiag}\left\{\boldsymbol{R}_{\boldsymbol{yy}}(i)\right\}\Big]\mathrm{bdiag}^{-1}\{\boldsymbol{R}_{\boldsymbol{yy}}(i)\}\right\}\,.\end{aligned}\tag{1.35}$$

Comparing Eq. 1.35 to the HOS-SIRP update (Eq. 1.30) shows that due to the fact that only SOS are utilized, we obtain the same update with the nonlinearity (Eq. 1.28) omitted, i.e., $\phi_{y_q,D}(u_q(iL+j))=1$, $q=1,\ldots,P$. Therefore, the SOS natural gradient update also exhibits the inherent normalization by the auto-correlation matrices which leads to very robust convergence behaviour in real-world environments. Moreover, due to the inversion of channel-wise $D\times D$ submatrices, $N>D$ instead of $N>PD$ is again sufficient for the estimation of the correlation matrices.

In Fig. 1.4 the structure of the cost function in the case of SOS and idealized/simplified mechanism of the adaptation update (Eq. 1.35) is illustrated. By assuming the multivariate Gaussian PDF (Eq. 1.33) and then minimizing $\mathcal{J}(m,\boldsymbol{W})$, all cross-correlations for D time-lags are reduced and thus the algorithm exploits nonwhiteness. Nonstationarity is utilized by minimizing the correlation matrices simultaneously for several blocks i. Ideally, the cross-correlations will be equal to zero upon convergence which causes the update term to be zero because then $\boldsymbol{R}_{\boldsymbol{yy}}(i)-\mathrm{bdiag}\left\{\boldsymbol{R}_{\boldsymbol{yy}}(i)\right\}=\mathbf{0}$.

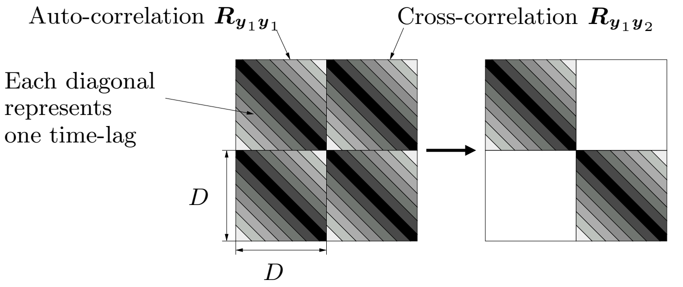

Auto-correlation $\boldsymbol{R}_{\boldsymbol{y}_1\boldsymbol{y}_1}$
Cross-correlation $\boldsymbol{R}_{\boldsymbol{y}_1\boldsymbol{y}_2}$
Each diagonal
represents
one time-lag
D
D

Fig. 1.4. Illustration of the diagonalization of the correlation matrices $\boldsymbol{R}_{\boldsymbol{yy}}(i)$ performed by the natural gradient update (Eq. 1.35) for the 2×2 case.

An alternative derivation of a SOS BSS algorithm leading to the same natural gradient update as given in Eq. 1.35 was presented in [17]. There, the derivation was based on a generalized version of the cost function used in [64], which also simultaneously exploits nonwhiteness and nonstationarity of the sources.

Covariance method
Correlation method
Toeplitz structure

1.2.4 Estimation of the Correlation Matrices and an Efficient Normalization Strategy

In this section some implementation aspects are addressed which allow to reduce the computational complexity. The first aspect is the block-based estimation of the short-time output correlation matrices $\boldsymbol{R}_{\boldsymbol{y}_p\boldsymbol{y}_q}(i)$ for nonstationary signals for which two basic methods exist: The so-called *covariance method* and the *correlation method* as they are known from linear prediction problems [58]. It should be emphasized that the terms covariance method and correlation method are not based upon the standard usage of the covariance function as the correlation function with the means removed. In the definition of the correlation matrices in Eq. 1.23 the more accurate covariance method was introduced. To obtain more efficient implementations, the computationally less complex correlation method can be used which is obtained by assuming stationarity within each block i. This leads to a Toeplitz structure of the matrices $\boldsymbol{R}_{\boldsymbol{y}_p\boldsymbol{y}_q}(i)$ and thus simplifies the computation of the matrix products [6]. Furthermore, it is important to note that regardless of the estimation of the correlation matrices, the matrix product of Sylvester matrices $\boldsymbol{W}_{pq}$ and the remaining matrices in the update Eqs. 1.30 and 1.35 can be described by linear convolutions due to the Sylvester structure involved.

As a second aspect, we discuss the inherent normalization by the auto-correlation matrices in Eqs. 1.30 and 1.35 which is introduced by the usage of multivariate PDFs as pointed out in the previous section. The normalization is desirable as it guarantees fast convergence of the adaptive filters even for large filter lengths and correlated input signals. On the other hand this poses the problem of large computational complexity due to the required matrix inversion of P matrices of size $D \times D$. The complexity of a straightforward implementation is $\mathcal{O}(D^3)$ for using the covariance method and $\mathcal{O}(D^2)$ for the correlation method due to the Toeplitz structure involved. However, as D may be even larger than 1000 for realistic environments this is still prohibitive for a real-time implementation on regular PC platforms. Therefore, approximations are desirable which reduce the complexity with minimum degradation of the separation performance.

One possible solution is to approximate the auto-correlation matrices $\boldsymbol{R}_{\boldsymbol{y}_q\boldsymbol{y}_q}(i)$ by a diagonal matrix, i.e., by the output signal powers

$$\boldsymbol{R}_{\boldsymbol{y}_q\boldsymbol{y}_q}(i) \approx \frac{1}{N}\sum_{j=0}^{N-1} \operatorname{diag}\left\{\boldsymbol{y}_q(iL+j)\boldsymbol{y}_q^{\mathrm{T}}(iL+j)\right\}. \tag{1.36}$$

for $q = 1, \ldots, P$, where the operator diag$\{\boldsymbol{A}\}$ sets all off-diagonal elements of matrix $\boldsymbol{A}$ to zero. This approximation is comparable to the one in the well-known normalized least mean squares (NLMS) algorithm in supervised adaptive filtering approximating the RLS algorithm [41]. It should be noted that the SOS natural gradient algorithm based on Eq. 1.35 together with the approximation 1.36 was also heuristically introduced for the case $D = L$

Regularization
Blind source
separation!instantaneous
Blind source
separation!narrowband

in [2, 70] as an extension of [50] incorporating several time-lags. It should be pointed out that also a more sophisticated approximation of the normalization is possible. One approach which exploits the efficiency of computations in the DFT domain is outlined in the next section.

For blocks with speech pauses and low background noise the normalization by the auto-correlation matrix $\boldsymbol{R}_{\boldsymbol{y}_q\boldsymbol{y}_q}$ leads to the inversion of an ill-conditioned matrix or in the case of the approximation (Eq. 1.36) to a division by very small output powers or even by zero becomes likely and thus, the estimation of the filter coefficients becomes very sensitive. For a robust adaptation $\boldsymbol{R}_{\boldsymbol{y}_q\boldsymbol{y}_q}$ is replaced by a regularized version $\boldsymbol{R}_{\boldsymbol{y}_q\boldsymbol{y}_q} + \delta_{y_q}\boldsymbol{I}$. The basic feature of the regularization is a compromise between fidelity to data and fidelity to prior information about the solution [23]. As the latter increases robustness but leads to biased solutions, similarly to supervised adaptive filtering [19], a dynamical regularization

$$\delta_{y_q} = \delta_{\max} e^{-\sigma_{y_q}^2/\sigma_0^2} \tag{1.37}$$

can be used with two parameters $\delta_{\max}$ and σ_0^2. This exponential method provides a smooth transition between regularization for low output power $\sigma_{y_q}^2$ and data fidelity whenever the output power is large enough. Other popular strategies are the fixed regularization which simply adds a constant value to the output power $\delta_{y_q} = \text{const}$ and the approach of choosing the maximum out of the output signal power $\sigma_{y_q}^2$ and a fixed threshold δ_{th}.

1.2.5 On Broadband and Narrowband BSS Algorithms in the DFT Domain

In the previous sections it was shown how different time-domain algorithms can be derived from the TRINICON framework. On the other hand, for convolutive mixtures the classical approach of frequency-domain BSS appears to be an attractive alternative because all techniques originally developed for instantaneous BSS can typically be applied independently in each frequency bin, e.g., [48]. Unfortunately, this traditional narrowband approach exhibits several limitations as identified in, e.g., [10,55,75]. In particular, the permutation problem pointed out in Sec. 1.1, which is inherent in BSS may then also appear independently in each frequency bin so that extra repair measures have to be taken to address this *internal* permutation. Moreover, problems caused by circular convolution effects due to the narrowband approximation are reported in, e.g., [75].

To exploit the computational efficiency it is desirable to derive approaches in the DFT domain, but on the other hand the above-mentioned problems of the narrowband approach should be avoided. This can be achieved by transforming the equations of the TRINICON framework into the DFT domain in a rigorous way (i.e., without any approximations) as was shown in [16, 17]. As in the case of time-domain algorithms, the resulting generic DFT-domain

Blind source separation!broadband
Multivariate probability density function
Multivariate score function
TRINICON

broadband BSS may serve both as a unifying framework for existing algorithms, and also as a starting point for developing new improved algorithms by a considerate choice of *selective* approximations as shown in, e.g., [7,16]. Fig. 1.5 gives an overview on the most important classes of DFT-domain BSS algorithms known so far (various more special cases may be developed in the future). A very important observation from this framework using multivariate PDFs is that the internal permutation problem is avoided. This is achieved by the following two elements:

1. Constraint matrices (consisting of an inverse DFT followed by a zeroing of several elements in the time domain and a subsequent DFT) appear in the generic DFT-domain formulation (see, e.g., [16,17]) and describe the inter-frequency correlation between DFT components.
2. The coupling between the DFT bins is additionally ensured by the multivariate score function which is derived from the multivariate PDF [16]. As an example, for SIRPs the argument of the multivariate score function (which is in general a nonlinear function) is $\boldsymbol{y}_q^{\mathrm{T}}(iL+j)\boldsymbol{R}^{-1}_{\boldsymbol{y}_q\boldsymbol{y}_q}(i)\boldsymbol{y}_q(iL+j)$ according to Eq. 1.22. Even for the simple case $\boldsymbol{R}^{-1}_{\boldsymbol{y}_q\boldsymbol{y}_q}(i) = \boldsymbol{I}$, where we have $\boldsymbol{y}_q^{\mathrm{T}}(iL+j)\boldsymbol{y}_q(iL+j) = \|\boldsymbol{y}_q(iL+j)\|^2$, i.e., the quadratic norm, and – due to the Parseval theorem – the same in the DFT domain, i.e., the quadratic norm over all DFT components ensures a coupling between all DFT bins. From this we immediately see that for the adaptation of an individual DFT bin all DFT bins are taken into account simultaneously so that the internal permutation problem is at least mitigated if not completely avoided.

This illustrates that the dependencies among all DFT components (including higher-order dependencies) are inherently taken into account in the TRINICON framework. The traditional narrowband approach (with the internal permutation problem) would result as a special case if we assume all DFT components to be statistically independent from each other which is of course not the case for real-world broadband signals such as speech and audio signals. Actually, in the traditional narrowband approach, the additionally required repair mechanisms for permutation alignment try to exploit such inter-frequency dependencies [51].

In Fig. 1.5 it can be seen that the TRINICON framework allows to introduce several *selective* approximations which can be used to cover several well-known algorithms and also to derive new algorithms. Among these algorithms, the system described in [7] has turned out to be very efficient. There, only the normalization by the auto-correlation matrix in the SOS BSS update (Eq. 1.35) has been approximated by a narrowband inverse which allows to perform for each channel a scalar inversion for each DFT bin instead of a $D \times D$ matrix inverse. Therefore, this algorithm is included in the experimental evaluation in the next section. More details and a pseudo-code of the algorithm can be found in [7].

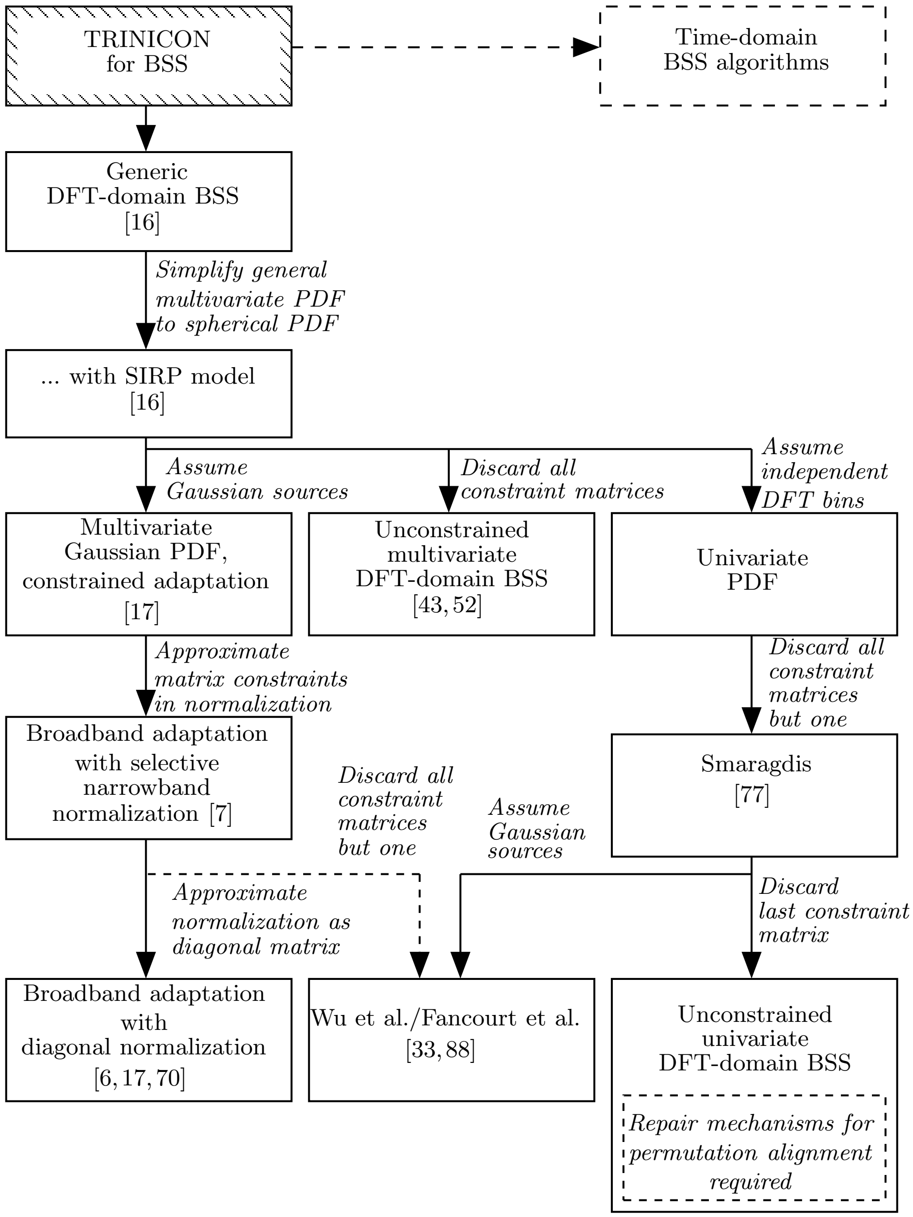


Fig. 1.5. Overview of BSS algorithms in the DFT domain.

1.2.6 Experimental Results for Reverberant Environments

The separation performance of various BSS algorithms derived from the TRINICON framework is shown for a living room scenario with a reverberation time $T_{60} = 200\,\mathrm{ms}$. Two sources have been placed at a distance of 1 m at $-20°$, $40°$ from a microphone pair with omnidirectional sensors and the signals have been sampled at $f_\mathrm{s} = 16\,\mathrm{kHz}$. To cope well with the reverberation, the demixing filter length has been chosen to $L = 1024$ taps. The nonwhiteness is exploited by the memory of $D = L = 1024$ introduced in the multivariate PDFs and in the correlation matrices. For accurate estimation of these quantities a block length of $N = 2048$ was chosen. For all examined algorithms the Sylvester constraint $\mathcal{SC}_\mathcal{R}$ together with the initialization $w_{pp,15} = 1$, $p = 1, 2$ (all other taps are set to zero) is used due to the increased versatility [4,6] and the correlation method is used for the estimation of the correlation matrices to reduce computational complexity. For the iterative adaptation procedure the block-online update with $\ell_\mathrm{max} = 5$ offline iterations together with an adaptive stepsize is used (for details, see [6]). This allows online processing of the sensor signals and fast convergence when iterating ℓ_max times on the same data block.

The evaluated algorithms include the computationally complex second-order statistics algorithm (Eq. 1.35) as well as efficient algorithms obtained by applying several approximations to the generic algorithm. A list of all algorithms is given in Tab. 1.1. The performance of the algorithms is measured in terms of the segmental signal-to-interference ratio (SIR) improvement ΔSIR_seg. The segmental SIR measures the ratio of the power of the desired signal versus the power of the interfering signals and then averages this quantity over all P channels.

Table 1.1. List of algorithms evaluated in the reverberant living room scenario.

Identifier	Algorithmic description
(A)	Broadband SOS algorithm (Eq. 1.35) based on the multivariate Gaussian PDF.
(B)	Broadband SOS algorithm based on multivariate Gaussian PDF (Eq. 1.35) with normalization approximated as a narrowband inverse [7].
(C)	Broadband SOS algorithm based on multivariate Gaussian PDF (Eq. 1.35) with normalization approximated as a scaling by the output signal variance (Eq. 1.36) [6, 17].
(D)	Narrowband SOS algorithm based on multivariate Gaussian PDF where the coupling between the DFT bins is ensured by one remaining constraint matrix [17].

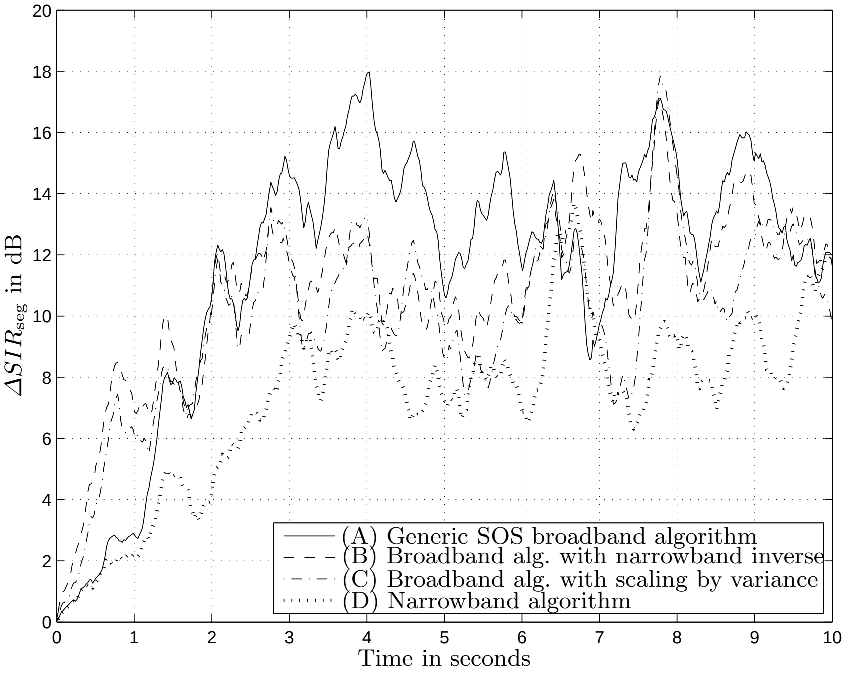


Fig. 1.6. Segmental SIR improvement $\Delta\overline{SIR}_{\mathrm{seg}}$ for the second-order statistics algorithms (A)-(D) evaluated in the living room scenario ($T_{60} = 200\,\mathrm{ms}$) for two source position setups.

The experimental results in Fig. 1.6 show that the SOS algorithm (A) provides the best performance. However, its high computational complexity prevents a real-time implementation on current state-of-the-art hardware platforms for such large demixing filter lengths. Therefore approximations are needed which minimally affect the separation performance but result in computationally efficient algorithms. As pointed out above, the main complexity in the second-order statistics algorithms is caused by the inverse of the autocorrelation matrix for each output channel. This inverse is approximated in the broadband algorithm (B) by a narrowband inverse which leads to a scalar inversion in each DFT bin [7]. The algorithm (B) can be implemented in real-time on regular PC hardware and it can be seen in Fig. 1.6 that the separation performance is only slightly reduced. In the broadband algorithm (C) the normalization is further simplified by using the variance of each output signal [6,17] as shown in Sec. 1.2.4. This means that the normalization is not frequency-dependent anymore. In the narrowband algorithm (D) all constraint matrices except one are approximated [17]. This means that the narrowband normalization is done analogously to algorithm (B), however, due to discarding all constraint matrices except for one, the complete decoupling of the

DFT bins is only prevented by the last remaining constraint matrix. Thus, the algorithm (D) already suffers from the permutation and scaling problem occuring in each DFT bin. This explains the inferior separation performance compared to the broadband algorithms (A)-(C). More extensive simulations for algorithms derived from the TRINICON framework including different source positions and reverberation times can be found in [8].

1.3 Extensions for Blind Source Separation in Noisy Environments

In the previous section only noiseless reverberant environments were considered with the maximum number of simultaneously active point sources Q assumed to be equal to the number of sensors P. However, in realistic scenarios, in addition to the point sources to be separated, some background noise will usually be present. Thus, in general BSS faces two different challenges in noisy environments:

1. The *adaptation* of the demixing BSS filters should be *robust* to the *noise signals* $n_1(n), \dots, n_P(n)$ to ensure high separation performance of the desired point sources $s_1(n), \dots, s_P(n)$. This means that the signal-to-interference ratio (SIR) should not deteriorate compared to the noiseless case.
2. The *noise* contribution contained in the separated BSS output signals *should be suppressed*, i.e., the signal-to-noise ratio (SNR) should be maximized.

Both requirements must be met if BSS should be attractive for noisy environments.

According to the literature (see e.g., [25, 48] and references therein), it has been tried to address the first point by developing noise-robust BSS algorithms. However, so far this has been considered only for the instantaneous BSS case. Additionally, several assumptions such as spatial or temporal uncorrelatedness are usually imposed on the noise signals allowing to generate optimization criteria which are not affected by the noise signals. However, in the case of convolutive BSS for acoustic signals these assumptions for the noise signals are too restrictive. In the next section when discussing a model for background noise, we will see that in realistic scenarios the background noise at the sensors is *temporally correlated* and for low frequencies and/or small microphone spacings it will also be *spatially correlated*. This does not allow the application of the noise-robust instantaneous BSS algorithms presented in the literature to the noisy convolutive BSS problem.

A more promising approach for increasing the robustness of the BSS adaptation are pre-processing methods. In Sec. 1.3.2 we will describe single-channel and multi-channel methods in order to remove the bias of the second-order

Diffuse sound field

Magnitude-squared coherence

correlation matrices caused by the noise. This will lead to a better performance of previously discussed BSS algorithms.

Another approach to the noisy convolutive BSS problem is the application of post-processing methods to the outputs of the BSS system. Without preprocessing the separation performance of the BSS algorithms will decrease in noisy environments. Therefore, the post-processing technique has to aim at the suppression of both, background noise and residual crosstalk from interfering point sources which could not be cancelled by the BSS demixing filters. This will be discussed in detail in Sec. 1.3.3.

1.3.1 Model for Background Noise in Realistic Environments

A model often used to describe background noise is the 3-dimensional isotropic sound field which is also termed diffuse sound field [56]. It can be modeled by an infinite number of statistically independent point sources which are uniformly distributed on a sphere. The phases of the emitted background noise signals are uniformly distributed between 0 and 2π. If the radius of the sphere is $r \to \infty$, then the propagating waves from each point source picked up be the microphones x_p can be assumed to be plane waves.

The diffuse sound field allows to describe, e.g., speech babble noise in a cafeteria, which is generated by a large number of background speakers or exterior noise recorded in the passenger compartment of a car which is a superposition of many different sources such as, e.g., motor, wind, or street noise. Moreover, the diffuse sound field is also often used to model reverberation [56]. This requires that the direct sound and the reflections are assumed to be mutually incoherent, i.e., the phase relations between the sound waves are neglected and thus, a superposition of the sound waves only results in a summation of the sound intensities. As the convolutive BSS demixing system accounts for the phase relations by the FIR filters of length L only the reflections exceeding the time-delay covered by L filter taps can be considered as being of diffuse nature. This case applies to highly reverberant environments such as, e.g., lecture rooms, or train stations.

In the convolutive BSS model depicted in Fig. 1.1 we assumed that the number of simultaneously active point sources Q is less or equal to the number of sensors P. Due to the limited number of point sources Q in the BSS scenario, we thus cannot model the diffuse sound field by an infinite number of point sources. Therefore, they are included in the BSS model in Fig. 1.1 as noise components $n_p(n)$, $p = 1, \ldots, P$ which are additively mixed to each microphone signal $x_p(n)$.

An adequate quantity to classify the sound field at the sensors is the magnitude-squared coherence (MSC) function whose estimate for the m-th data block and ν-th DFT bin is given by

$$\left|\underline{\Gamma}^{(\nu)}_{x_1x_2}(m)\right|^2 = \frac{\left|\underline{S}^{(\nu)}_{x_1x_2}(m)\right|^2}{\underline{S}^{(\nu)}_{x_1x_1}(m)\underline{S}^{(\nu)}_{x_2x_2}(m)}. \tag{1.38}$$

The estimation of the power-spectral densities $\underline{S}^{(\nu)}_{x_p x_q}(m)$ for nonstationary signals such as speech is usually performed using recursive averaging with a forgetting factor γ given as

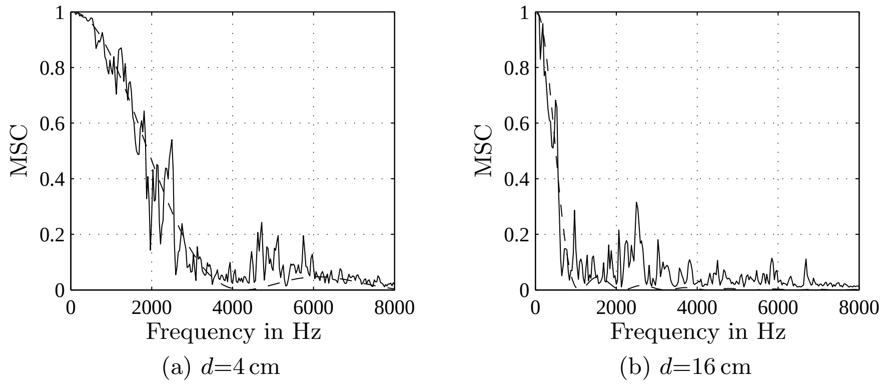
$$\underline{S}^{(\nu)}_{x_p x_q}(m) = \gamma \underline{S}^{(\nu)}_{x_p x_q}(m-1) + (1-\gamma)\underline{X}^{(\nu)}_p(m)\underline{X}^{(\nu)^*}_q(m). \tag{1.39}$$

A long-term estimate of the MSC can be obtained by averaging the short-term MSC $|\underline{\Gamma}^{(\nu)}_{x_1 x_2}(m)|^2$ over all blocks.

In an ideally diffuse sound field the MSC between the microphone signals $x_1(n)$ and $x_2(n)$ is given by

$$\left|\underline{\Gamma}^{(\nu)}_{x_1 x_2}\right|^2 = \frac{\sin^2\left(2\pi\nu R^{-1} f_{\mathrm{s}} d\, c^{-1}\right)}{\left(2\pi\nu R^{-1} f_{\mathrm{s}} d\, c^{-1}\right)^2}, \tag{1.40}$$

where d denotes the distance between the microphones and R is the DFT length. This result assumes omnidirectional sensor characteristics and was first presented in [26] (a detailed derivation can be found, e.g., in [8, 60]). Eq. 1.40 reflects that the noise components $n_p(n)$ which originated from a diffuse sound field are strongly correlated between the sensors at low frequencies but less correlated for higher frequencies. Additionally, each $n_p(n)$, $p = 1, \ldots, P$ may also contain sensor noise which is usually assumed independent across different sensors. For comparison, the MSC for a point source $s_q(n)$ is equal to one. In Fig. 1.7 the estimated MSC of car noise is shown. A two-element omnidirectional microphone array was positioned in the passenger compartment at the interior mirror and two different spacings of $d = 4\,\mathrm{cm}$ and $d = 16\,\mathrm{cm}$ have been examined. The car noise was measured while driving through a suburban area. The estimate of the MSC (Eq. 1.38) was obtained by using the recursive averaging procedure (Eq. 1.39) with $\gamma = 0.9$, DFT length $R = 512$, and using Hann windowing. The long-term estimate $|\Gamma^{(\nu)}_{x_1 x_2}|^2$ was



(a) d=4 cm (b) d=16 cm

Fig. 1.7. MSC $|\Gamma^{(\nu)}_{x_1 x_2}|^2$ of car noise measured at two sensors $x_1(n)$ and $x_2(n)$ positioned at the interior mirror in a car compartment for different sensor spacings d.

Pre-processing
Bias removal

calculated by averaging over all blocks for a signal length of 20 sec. It can be seen that the MSC of the measured data (solid) corresponds very well to the $\sin^2(x)/x^2$ characteristic of the MSC of an ideal diffuse sound field (dashed) for both microphone spacings. Therefore, it can be concluded that the MSC of car noise can be approximated by the MSC of a diffuse sound field. Additionally, in [62] it was shown experimentally that also office noise originating from computer fans and hard disk drives can be assumed to exhibit the MSC of a diffuse noise field.

1.3.2 Pre-Processing for Noise-Robust Adaptation

From the literature only few pre-processing approaches for BSS in noisy environments are known. If the number of sensors P is equal to the number of sources Q, as considered in this chapter, then usually so-called *bias removal techniques* are used which aim at estimating and subtracting the contribution of the noise in the sensor signal itself or in the second-order correlation matrix and possibly also in the higher-order relation matrix of the sensor signals. These techniques will be discussed in the following. If more sensors than sources are available, i.e., $P > Q$, then also *subspace techniques* can be used as a pre-processing step to achieve a suppression of the background noise. As we restricted ourselves in this chapter to the case $P = Q$ the subspace approaches will not be treated here, but a summary and an outline of possible directions of future research can be found in [8].

The signal model in matrix-vector notation (Eq. 1.7) yields the BSS output signals $\boldsymbol{y}(n)$ containing D output signal samples for each of the $Q = P$ channels. If background noise $n_p(n)$ is superimposed at each sensor $p = 1, \ldots, P$, the signal model can be decomposed as

$$\begin{aligned} \boldsymbol{y}(n) &= \boldsymbol{W}^{\mathrm{T}} \boldsymbol{x}(n) \\ &= \boldsymbol{W}^{\mathrm{T}} \left(\boldsymbol{H}^{\mathrm{T}} \boldsymbol{s}(n) + \boldsymbol{n}(n) \right) \end{aligned} \tag{1.41}$$

where the background noise and speech samples are contained in the column vectors

$$\boldsymbol{s}(n) = \left[\boldsymbol{s}_1^{\mathrm{T}}(n), \ldots, \boldsymbol{s}_P^{\mathrm{T}}(n)\right]^{\mathrm{T}}, \tag{1.42}$$

$$\boldsymbol{s}_p(n) = \left[s_p(n), \ldots, s_p(n - 2L - M + 2)\right]^{\mathrm{T}}, \tag{1.43}$$

$$\boldsymbol{n}(n) = \left[\boldsymbol{n}_1^{\mathrm{T}}(n), \ldots, \boldsymbol{n}_P^{\mathrm{T}}(n)\right]^{\mathrm{T}}, \tag{1.44}$$

$$\boldsymbol{n}_p(n) = \left[n_p(n), \ldots, n_p(n - 2L + 1)\right]^{\mathrm{T}}, \tag{1.45}$$

and the matrix $\boldsymbol{H}$ is composed of channel-wise Sylvester matrices $\boldsymbol{H}_{qp}$ of size $(M + 2L - 1) \times 2L$ containing the mixing FIR filters $h_{qp,\kappa}$, $\kappa = 0, \ldots, M - 1$. It can be seen from the noisy signal model (Eq. 1.41) that the second-order correlation matrix $\boldsymbol{R}_{\boldsymbol{yy}}(n)$ and also the higher-order relation matrix $\boldsymbol{R}_{\boldsymbol{y}\phi(\boldsymbol{y})}(n)$

will contain a bias due to the background noise. Due to the central limit theorem, the distribution of the diffuse background noise can be assumed to be closer to a Gaussian than the distribution of the speech signals. Therefore, the bias will be larger for the estimation of the cross-correlation matrix $\boldsymbol{R}_{\boldsymbol{yy}}(n)$ and the background noise will affect the estimation of higher-order moments less. Therefore, we will focus on bias removal for second-order correlation matrices. The background noise $\boldsymbol{n}(n)$ and the point-source signals $\boldsymbol{s}(n)$ are assumed to be mutually uncorrelated so that the second-order correlation matrix $\boldsymbol{R}_{\boldsymbol{yy}}(n)$ with its channel-wise submatrices defined in Eq. 1.23 can be decomposed as

$$\boldsymbol{R}_{\boldsymbol{yy}}(n) = \boldsymbol{W}^{\mathrm{T}} \left(\boldsymbol{H}^{\mathrm{T}} \boldsymbol{R}_{\boldsymbol{ss}}(n) \boldsymbol{H} + \boldsymbol{R}_{\boldsymbol{nn}}(n) \right) \boldsymbol{W} \tag{1.46}$$

with the source correlation matrix $\boldsymbol{R}_{\boldsymbol{ss}}(n)$ and noise correlation matrix $\boldsymbol{R}_{\boldsymbol{nn}}(n)$ defined as

$$\boldsymbol{R}_{\boldsymbol{ss}}(n) = \frac{1}{N} \sum_{j=0}^{N-1} \boldsymbol{s}(n+j)\, \boldsymbol{s}^{\mathrm{T}}(n+j), \tag{1.47}$$

$$\boldsymbol{R}_{\boldsymbol{nn}}(n) = \frac{1}{N} \sum_{j=0}^{N-1} \boldsymbol{n}(n+j) \boldsymbol{n}^{\mathrm{T}}(n+j). \tag{1.48}$$

To remove the bias introduced by the background noise it is possible to either aim at estimating and subsequently removing the noise component in Eq. 1.41, e.g., by using single-channel noise reduction techniques, or to estimate and remove the noise correlation matrix $\boldsymbol{R}_{\boldsymbol{nn}}(n)$. The latter approach is already known from the literature on instantaneous BSS. There, usually spatially and temporally uncorrelated Gaussian noise is assumed, i.e., $\boldsymbol{R}_{\boldsymbol{nn}}(n)$ is a diagonal matrix (see, e.g., [24,25,30,48]). Moreover, most approaches assume that $\boldsymbol{R}_{\boldsymbol{nn}}(n)$ is known a-priori and stationary. However, in realistic scenarios usually temporally correlated background noise is present at the sensors. This noise can often be described by a diffuse sound field, leading to noise signals which are also spatially correlated for low frequencies (see Sec. 1.3.1). Additionally, background noise is in general nonstationary and its stochastic properties can at best be assumed slowly time-variant which thus requires a continuous estimation of the correlation matrix $\boldsymbol{R}_{\boldsymbol{nn}}(n)$ based on short-time stationarity according to Eq. 1.48. The following bias removal techniques, aiming at the noise signal $\boldsymbol{n}(n)$ or the noise correlation matrix $\boldsymbol{R}_{\boldsymbol{nn}}(n)$, will be examined under these conditions.

1.3.2.1 Single-Channel Noise Reduction

If the estimation and suppression of the noise components $\boldsymbol{n}_p(n)$ is desired for each sensor signal $\boldsymbol{x}_p(n)$ individually, then for each channel $p = 1, \ldots, P$ a single-channel noise reduction algorithm can be used. The estimation and suppression of background noise using one channel is already a long-standing research topic. In general, all algorithms consist of two main building blocks:

- the estimation of the noise contribution and
- the computation of a weighting rule to suppress the noise and enhance the desired signal.

An overview of various methods can be found, e.g., in [39].

In all well-known noise estimation methods in the literature usually the noise power spectral density (PSD) is estimated without recovering the phase of the clean signal but using the phase of the noisy signal instead. This is motivated by the fact that the human perception of speech is not much affected by a modification of the phase of the clean signal [83]. However, for BSS algorithms the relative phase of the signals acquired by different microphones is crucial as this information is implicitly used to suppress signals depending on their different directions of arrival. To evaluate the importance of amplitude and phase for pre-processing techniques applied to BSS algorithms, we will in the following generate pre-processed sensor signals by using the DFT-domain amplitude of the clean mixture signals and the phase of the noisy mixture signals. This corresponds to an optimum single-channel speech enhancement algorithm which perfectly estimates the amplitude of the clean mixture signal and thus, suppresses the background noise completely. These signals are then used as inputs for the second-order statistics BSS algorithm described in [7].

For this experiment we use two noisy scenarios. The first one is a car environment where a pair of omnidirectional microphones with a spacing of 20 cm was mounted to the interior mirror. The long-term SNR was adjusted to 0 dB which is a realistic value commonly encountered inside car compartments. Analogously to the BSS experiments in Sec. 1.2.6 a male and a female speech signal were convolved with the acoustic impulse response measured for the driver and co-driver positions. The second scenario corresponds to the cocktail party problem which is usually described by the task of listening to one desired point source in the presence of speech babble noise consisting of the utterances of many other speakers. The long-term statistics of speech babble are well described by a diffuse sound field, however, there may also be several other distinct noise point sources present. In our experiments we simulated such a cocktail party scenario inside a living room environment where speech babble noise was generated by a circular loudspeaker array with a diameter of 3 m. The two omnidirectional microphones with a spacing of 20 cm were placed in the center of the loudspeaker array from which 16 speech signals were reproduced to simulate the speech babble noise. Additionally, two distinct point sources at a distance of 1 m and at the angles of 0° and $-80°$ were used to simulate the desired and one interfering point source, respectively. The long-term input SNR at the microphones has been adjusted for the living room scenario to 10 dB. This is realistic, as due to the speech-like spectrum of the background noise the microphone signals exhibiting higher SNR values are perceptually already as annoying as those with significantly lower SNR values for lowpass car noise.

Due to the perfect estimation of the clean signal amplitude the background noise is almost inaudible. However, the results in Fig. 1.8 show that due to the noisy phase for the car environment no improvement in terms of separation performance can be obtained. Similarly, for the cocktail party scenario only a small improvement in terms of separation of the point sources can be achieved. Further experiments also indicated that when using a realistic state-of-the-art noise reduction algorithm as, e.g., proposed in [63], then also the improvements shown in Fig. 1.8(b) disappear. Therefore, it is concluded that pre-processing by single-channel noise reduction algorithms only suppresses the background noise, but does not improve the degraded separation

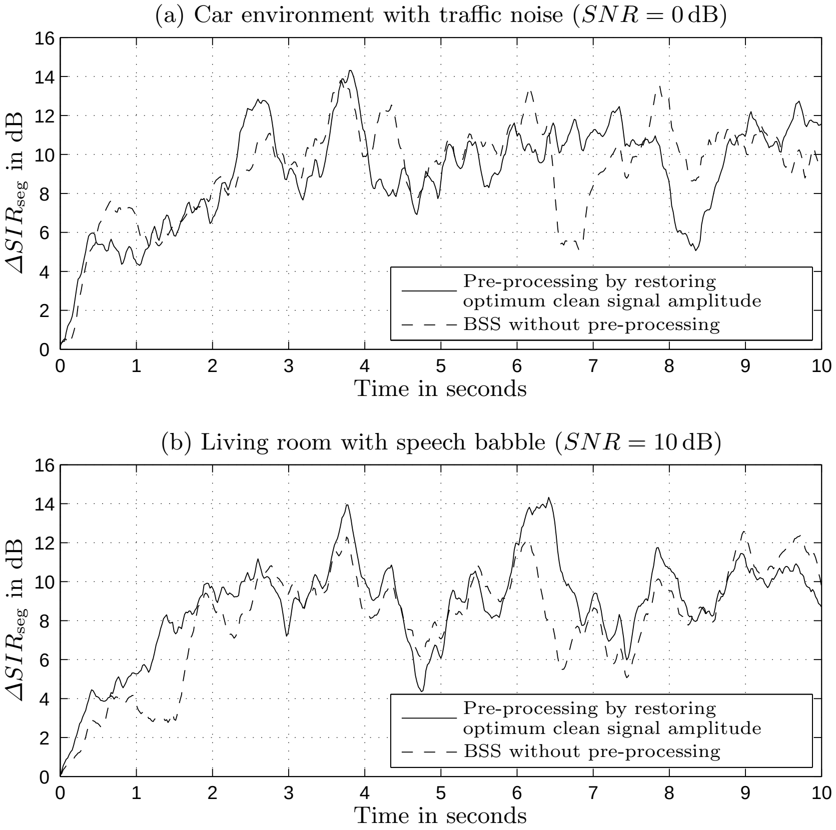


Fig. 1.8. Segmental SIR improvement ΔSIR_{seg} depicted over time for two noisy environments. Speech separation results are shown for the BSS outputs adapted with the noisy mixtures and for BSS with pre-processing by restoring the magnitude of the clean mixture signals but with the phase of the noisy mixtures.

performance of the subsequent BSS algorithm. To improve the separation, it is crucial that both, amplitude and phase of the clean mixture signals are estimated. This usually requires multi-channel methods as presented in the next section.

Bias removal!multi-channel
Blind source separation!instantaneous
Minimum statistics

1.3.2.2 Multi-Channel Bias Removal

To also account for the phase contribution of the background noise we first review briefly some methods initially proposed for instantaneous BSS which aim at estimating and subsequently removing the noise correlation matrix $\boldsymbol{R}_{\boldsymbol{nn}}(n)$. For convolutive BSS only a few approaches have been proposed so far: In [45] the special case of spatio-temporally white noise was addressed and has been extended to the diffuse noise case in [46]. There, stationarity of the noise was assumed and the preceding noise-only segments have been used for the estimation of the correlation matrix. Already earlier in [3, 6] a similar procedure was proposed where the minimum statistics approach [63] was used for the estimation of the noise characteristics. This method operates in the DFT domain and is based on the observation that the power of a noisy speech signal frequently decays to the power of the background noise. Hence by tracking the minima an estimate for the auto-power spectral density of the noise is obtained. However, due to the spatial correlation not only the auto- but also the cross-power spectral densities of the noisy signal $x_p(n)$ and the background noise $n_p(n)$ are required. They are estimated and averaged recursively for each DFT bin whenever we detect a minimum (i.e. speech pause) of the noisy speech signals. Thus, for slowly time-varying noise statistics this method gives an accurate estimate of the noise spectral density matrix used for the bias removal. In Fig. 1.9 the results of the approach in [3]

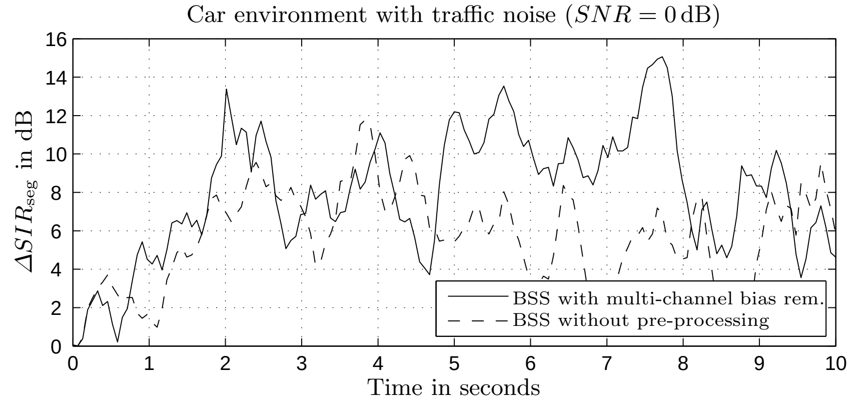


Fig. 1.9. Segmental SIR improvement ΔSIR_{seg} depicted over time for the noisy car environment. Speech separation results are shown for the BSS outputs adapted with the noisy mixtures and for BSS with pre-processing by multi-channel bias removal.

are shown in terms of the segmental SIR improvement for the separation of the two point sources. It can be seen that the pre-processing slightly improves the separation performance for the noisy car environment described in the previous section.

In the cocktail party scenario this approach did not achieve good results as the noise statistics is more time-variant and due to only few speech pauses of the point sources the noise PSD cannot be estimated very well.

In contrast to the single-channel bias removal techniques, the multi-channel approaches do not achieve any background noise reduction as they merely aim at providing a better estimate of the correlation matrix of the point sources which is then used for the adaptation of the demixing filter weights. To additionally suppress the background noise, this approach would have to be complemented by a post-processing technique. Note also that due to fewer speech pauses it is more difficult to estimate the noise correlation matrix for multiple active speakers compared to a single speaker as typically encountered in single-channel speech enhancement applications. Therefore, the estimation of the noise contribution may be done more reliably after the BSS stage where already a partial suppression of the interfering point sources is achieved. This will be investigated in detail for the post-processing approach discussed in Sec. 1.3.3

1.3.3 Post-Processing for Suppression of Residual Crosstalk and Background Noise

In Sec. 1.3.2 several pre-processing approaches have been discussed. It could be seen that for the case $P = Q$ only multi-channel bias removal methods achieved some noise robustness of the BSS algorithm. For these methods a reliable voice activity detection is crucial but might be difficult to realize in environments with several speech point sources so that in such cases post-processing methods are a preferable alternative. Post-processing methods have the advantage that the BSS system already achieves a suppression of the interfering point sources so that in each BSS output channel only some remaining interference of the other point sources is present. As will be shown later, this simplifies the estimation of the quantities required by the post-processing method. A suitable post-processing scheme is given by a single-channel postfilter $g_{q,\kappa}$ applied to each BSS output channel as shown in Fig. 1.10. The motivation of using a single-channel postfilter for each BSS output channel is twofold:

Firstly, it is desired that the remaining background noise is reduced at the BSS output channels. In [20] it was shown that the optimum solution for BSS leads to blind MIMO identification and thus, the BSS demixing system can be interpreted for each output channel as a blind adaptive interference canceller aiming at the suppression of the interfering point sources. As the background noise is usually described by a diffuse sound field, the BSS system achieves only limited noise suppression. However, from adaptive beamforming

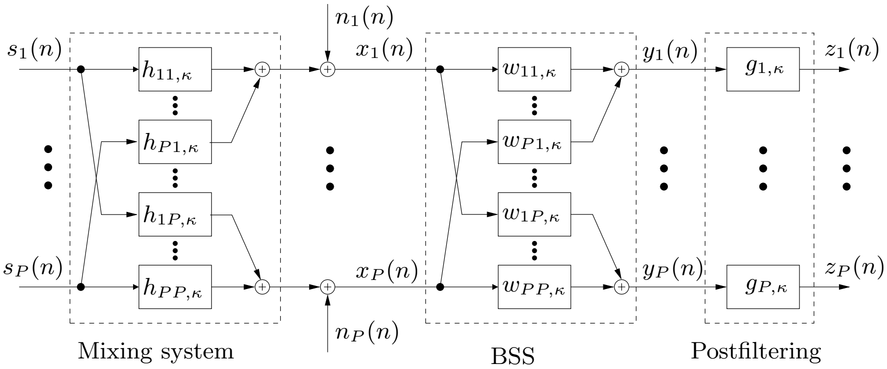


Fig. 1.10. Noisy BSS model combined with postfiltering.

(e.g., [76]) it is known that in such environments the concatenation of an adaptive interference canceller with a single-channel postfilter can improve the noise reduction.

Secondly, BSS algorithms are in noisy environments usually not able to converge to the optimum solution due to the bias introduced by the background noise. Moreover, moving point sources or an insufficient demixing filter length, which only partly covers the existing room reverberation, may lead to reduced signal separation performance and thus, to the presence of residual crosstalk from interfering point sources at the BSS output channels. In such situations, the single-channel postfilter should be designed such that it also provides additional separation performance. Analogously, similar considerations have led to a single-channel postfilter in acoustic echo cancellation which was first proposed in [11, 59].

The reduced separation quality due to an insufficient demixing filter length in realistic environments was the motivation of several single-channel postfilter approaches that have been previously proposed in the BSS literature [22,68,69,72,80,82]. Nevertheless, a comprehensive treatment of the simultaneous suppression of residual crosstalk and background noise is still missing and will be presented in the following sections. We will first discuss in Sec. 1.3.3.1 the advantages of the implementation of the single-channel postfilter in the DFT domain and will introduce a spectral gain function requiring the power spectral density (PSD) estimates of the residual crosstalk and background noise. Then, the signal model for the residual crosstalk and the background noise will be discussed in Sec. 1.3.3.2 allowing to point out the relationships to previous post-processing approaches. The chosen signal model will lead to the derivation of a novel residual crosstalk PSD estimation and additionally the estimation of the background noise will be addressed. Subsequently, experimental results will be presented which illustrate the improvements that can be obtained by the application of single-channel postfilters both, in terms of SIR and SNR.

1.3.3.1 Spectral Weighting Function for a Single-Channel Postfilter

The BSS output signals $y_q(n)$, $q = 1, \ldots, P$ can be decomposed for the q-th channel as

$$y_q(n) = y_{\mathrm{s}_r,q}(n) + y_{\mathrm{c},q}(n) + y_{\mathrm{n},q}(n), \tag{1.49}$$

where $y_{\mathrm{s}_r,q}(n)$ is the component containing the desired source $s_r(n)$. As a possible permutation of the separated sources at the BSS outputs, i.e., $r \neq q$ does not affect the post-processing approach we will simplify the notation and denote in the following the desired signal component in the q-th channel as $y_{\mathrm{s},q}(n)$. The quantity $y_{\mathrm{c},q}(n)$ is the residual crosstalk component from the remaining point sources that could not be suppressed by the BSS algorithm and $y_{\mathrm{n},q}(n)$ denotes the contribution of the background noise.

From single-channel speech enhancement (e.g., [12]) or from the literature on single-channel postfiltering for beamforming (e.g., [76]) it is well-known that it is beneficial to utilize the DFT-domain representation of the signals and estimate the single-channel postfilter in the DFT domain. Thus, N_{post} samples are combined to an output signal block which is, after applying a windowing operation, transformed by the DFT of length $R_{\mathrm{post}} \geq N_{\mathrm{post}}$ yielding the DFT-domain representation of the output signals as

$$\underline{Y}_q^{(\nu)}(m) = \underline{Y}_{\mathrm{s},q}^{(\nu)}(m) + \underline{Y}_{\mathrm{c},q}^{(\nu)}(m) + \underline{Y}_{\mathrm{n},q}^{(\nu)}(m) \tag{1.50}$$

where $\nu = 0, \ldots, R_{\mathrm{post}} - 1$ is the index of the DFT bin and m denotes the block time index. The advantage is that in the DFT domain speech signals are sparser, i.e., we can find regions in the time-frequency plane where the individual speech sources do not overlap (see e.g., [90]). This property is often exploited in underdetermined blind source separation where there are more simultaneously active sources than sensors (e.g., [29,84]). Here, this sparseness is used for the estimation of the quantities necessary for the implementation of the spectral gain function. A block diagram showing the main building blocks of a DFT-based postfilter is given in Fig. 1.11. There it can already be seen that analogously to single-channel speech enhancement or post-filtering applied to beamforming or acoustic echo cancellation, the DFT bins are treated in a narrowband manner as all computations are carried out independently in each DFT bin. Because of the narrowband treatment we have to ensure that circular convolution effects, appearing due to the signal modification by the spectral weighting, are not audible. Thus, the enhanced output signal z_q, which is the estimate $\hat{y}_{\mathrm{s},q}(n)$ of the clean desired source component, is computed by the means of an inverse DFT using a weighted overlap-add method including a tapered analysis and synthesis windows as suggested in [38]. This is in contrast to the BSS algorithms derived from the TRINICON framework where the linear convolution of the sensor signals with the estimated FIR demixing system is implemented without approximations equivalently in the DFT domain by the overlap-save method. In contrast to postfiltering, the selective narrowband approximations which are applied in the TRINICON

Wiener filter

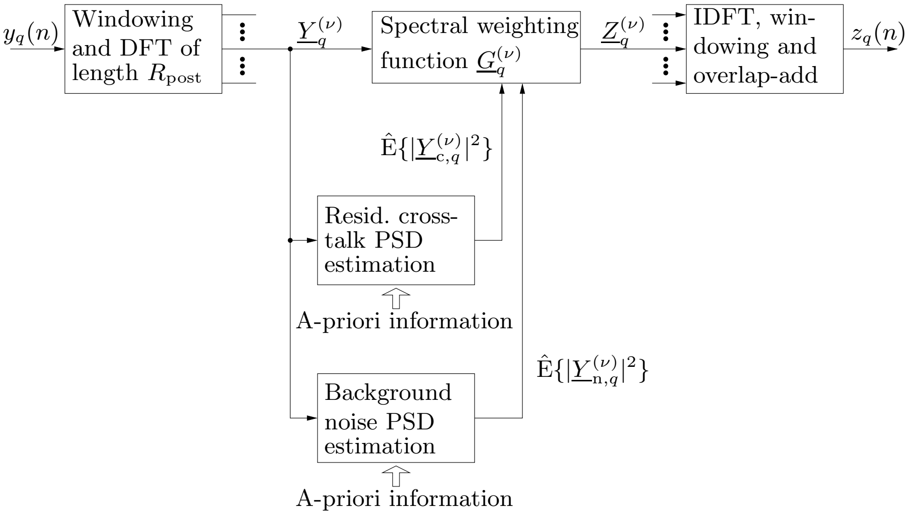

$y_q(n)$
Windowing and DFT of length R_{post}
$\underline{Y}_q^{(\nu)}$
Spectral weighting function $\underline{G}_q^{(\nu)}$
$\underline{Z}_q^{(\nu)}$
IDFT, windowing and overlap-add
$z_q(n)$
$\hat{\mathrm{E}}\{|\underline{Y}_{\mathrm{c},q}^{(\nu)}|^2\}$
Resid. crosstalk PSD estimation
A-priori information
$\hat{\mathrm{E}}\{|\underline{Y}_{\mathrm{n},q}^{(\nu)}|^2\}$
Background noise PSD estimation
A-priori information

Fig. 1.11. DFT-based single-channel postfiltering depicted for the ν-th DFT bin in the q-th channel.

framework and have been outlined in Sec. 1.2.5 have only been made in the adaptation process of the demixing filters to obtain efficient BSS algorithms.

According to Fig. 1.11 a spectral gain function $\underline{G}_q^{(\nu)}(m)$ in the ν-th DFT bin aiming at simultaneous suppression of residual crosstalk and background noise has to be derived. The output signal of the post-processing scheme is the estimate of the clean desired source signal

$$\underline{Z}_q^{(\nu)} = \hat{\underline{Y}}_{\mathrm{s},q}^{(\nu)} \tag{1.51}$$

and is given as

$$\underline{Z}_q^{(\nu)}(m) = \underline{G}_q^{(\nu)}(m)\underline{Y}_q^{(\nu)}(m)\,. \tag{1.52}$$

According to [5] we choose in this chapter to minimize the mean-squared error $\mathrm{E}\{(\underline{Z}_q^{(\nu)}(m) - \underline{Y}_{\mathrm{s},q}^{(\nu)}(m))^2\}$ with respect to $\underline{G}_q^{(\nu)}(m)$. This leads to the ν-th bin of the well-known Wiener filter for the q-th channel given as

$$\underline{G}_q^{(\nu)}(m) = \frac{\mathrm{E}\left\{\left|\underline{Y}_{\mathrm{s},q}^{(\nu)}(m)\right|^2\right\}}{\mathrm{E}\left\{\left|\underline{Y}_q^{(\nu)}(m)\right|^2\right\}}\,. \tag{1.53}$$

With the assumption that the desired signal component, the interfering signal components and the background noise in the q-th channel are all mutually uncorrelated, Eq. 1.53 can be expressed as

$$\underline{G}_q^{(\nu)}(m) = \frac{\mathrm{E}\left\{\left|\underline{Y}_{\mathrm{s},q}^{(\nu)}(m)\right|^2\right\}}{\mathrm{E}\left\{\left|\underline{Y}_{\mathrm{s},q}^{(\nu)}(m)\right|^2\right\} + \mathrm{E}\left\{\left|\underline{Y}_{\mathrm{c},q}^{(\nu)}(m)\right|^2\right\} + \mathrm{E}\left\{\left|\underline{Y}_{\mathrm{n},q}^{(\nu)}(m)\right|^2\right\}}\,. \tag{1.54}$$

From this equation it can be seen that for regions with desired signal *and* residual crosstalk or background noise components the output signal spectrum is reduced, whereas in regions without crosstalk or background noise the signal passed through. On the one hand this fulfills the requirement that an undisturbed desired source signal passes through the Wiener filter without any distortion. On the other hand, if crosstalk or noise is present, the magnitude spectrum of the noise or crosstalk attains a shape similar to that of the desired source signal, so that noise and crosstalk are therefore partially masked by the desired source signal. This effect was already exploited in postfiltering for acoustic echo cancellation aiming at the suppression of residual echo. There, this effect has been termed "echo shaping" [61]. Moreover, it can be observed in Eq. 1.54 that if the BSS system achieves the optimum solution, i.e., the residual crosstalk in the q-th channel $\underline{Y}_{\mathrm{c},q}^{(\nu)}(m) = 0$, then Eq. 1.54 reduces to the well-known Wiener filter for a signal with additive noise used in single-channel speech enhancement. To realize Eq. 1.54 in a practical system, the ensemble average $\mathrm{E}\{\cdot\}$ has to be estimated and thus, it is usually replaced by a time average $\hat{\mathrm{E}}\{\cdot\}$. Thereby, the Wiener filter is approximated by

$$\underline{G}_q^{(\nu)}(m) \approx \frac{\hat{\mathrm{E}}\Big\{\big|\underline{Y}_q^{(\nu)}(m)\big|^2\Big\} - \hat{\mathrm{E}}\Big\{\big|\underline{Y}_{\mathrm{c},q}^{(\nu)}(m)\big|^2\Big\} - \hat{\mathrm{E}}\Big\{\big|\underline{Y}_{\mathrm{n},q}^{(\nu)}(m)\big|^2\Big\}}{\hat{\mathrm{E}}\Big\{\big|\underline{Y}_q^{(\nu)}(m)\big|^2\Big\}}, \qquad (1.55)$$

where $\hat{\mathrm{E}}\{|\underline{Y}_q^{(\nu)}(m)|^2\}$, $\hat{\mathrm{E}}\{|\underline{Y}_{\mathrm{c},q}^{(\nu)}(m)|^2\}$, and $\hat{\mathrm{E}}\{|\underline{Y}_{\mathrm{n},q}^{(\nu)}(m|^2\}$ are the PSD estimates of the BSS output signal, residual crosstalk, and background noise, respectively. Due to the reformulation in Eq. 1.55 the unobservable desired signal PSD $\mathrm{E}\{|\underline{Y}_{\mathrm{s},q}^{(\nu)}(m)|^2\}$ does not have to be estimated. However, the main difficulty is still to obtain reliable estimates of the unobservable residual crosstalk and background noise PSDs. A novel method for this estimation process leading to high noise reduction with little signal distortion will be shown in the next section.

Moreover, an estimate of the observable BSS output signal PSD is required. The PSD estimates can be used to implement spectral weighting algorithms other than the Wiener filter as described, e.g., in [39].

1.3.3.2 Estimation of Residual Crosstalk and Background Noise

In this section a model for the residual crosstalk and background noise is introduced. Subsequently, based on the residual crosstalk model an estimation procedure will be given which relies on an adaptation control. Different adaptation control strategies will be outlined. Moreover, the estimation of the background noise PSD will be discussed.

Model of Residual Crosstalk and Background Noise

We restricted our scenario to the case that the number of microphones equals the maximum number of simultaneously active point sources. Therefore, the

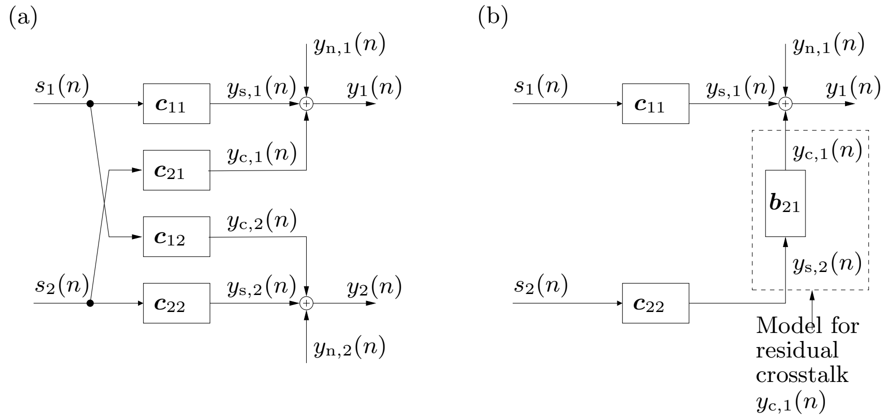


Fig. 1.12. (a) Representation of mixing and demixing system for the case $P = 2$ by using the overall system FIR filters $\boldsymbol{c}_{qr}$. (b) Resulting model for the residual crosstalk $y_{\mathrm{c},1}(n)$.

BSS algorithm is able to provide an estimate of one separated point source at each output $y_q(n)$. As pointed out above, due to movement of sources or long reverberation, the BSS algorithm might not converge fast enough to the optimum solution and thus, some residual crosstalk from point source interferers, denoted in the DFT domain by $\underline{Y}_{\mathrm{c},q}^{(\nu)}(m)$, remains in the BSS output. To obtain a good estimate of the residual crosstalk PSD $\mathrm{E}\{|\underline{Y}_{\mathrm{c},q}^{(\nu)}(m)|^2\}$ as needed for the post-filter in the q-th channel, we first need to set up an appropriate model.

In Fig. 1.12(a) the concatenation of the mixing and demixing systems are expressed by the overall filters $\boldsymbol{c}_{qr}$ of length $M + L - 1$ which denote the path from the q-th source to the r-th output. For simplicity, we have depicted the case $Q = P = 2$ in Fig. 1.12. As can be seen in Fig. 1.12(a), the crosstalk component $y_{\mathrm{c},1}(n)$ of the first output channel is determined in the case $Q = P = 2$ by the source signal $s_2(n)$ and the filter $\boldsymbol{c}_{21}$. However, as neither the original source signals nor the overall system matrix are observable, the crosstalk component $y_{\mathrm{c},1}(n)$ is expressed in Fig. 1.12(b) in terms of the desired source signal component $y_{\mathrm{s},2}(n)$ at the second output. This residual crosstalk model could be used if a good estimate of $y_{\mathrm{s},2}(n)$ is provided by the BSS system, i.e., if the source in the second channel is well-separated.

It should also be noted that even if $y_{\mathrm{s},2}(n)$ is available, then this model does not allow a perfect estimation of the residual crosstalk $y_{\mathrm{c},1}(n)$. This is due to the fact that for a perfect replica of $y_{\mathrm{c},1}(n)$ based on the input signal $y_{\mathrm{s},2}(n)$, the filter $\boldsymbol{b}_{21}$ has to model the combined system of $\boldsymbol{c}_{21}$ and the inverse of $\boldsymbol{c}_{22}$. However, $\boldsymbol{c}_{22}$ is in general a non-minimum phase FIR filter and thus, cannot be inverted in an exact manner by a single-input single-output system as was shown in [66]. Hence, analogously to single-channel blind dereverberation approaches, it is only possible to obtain an optimum

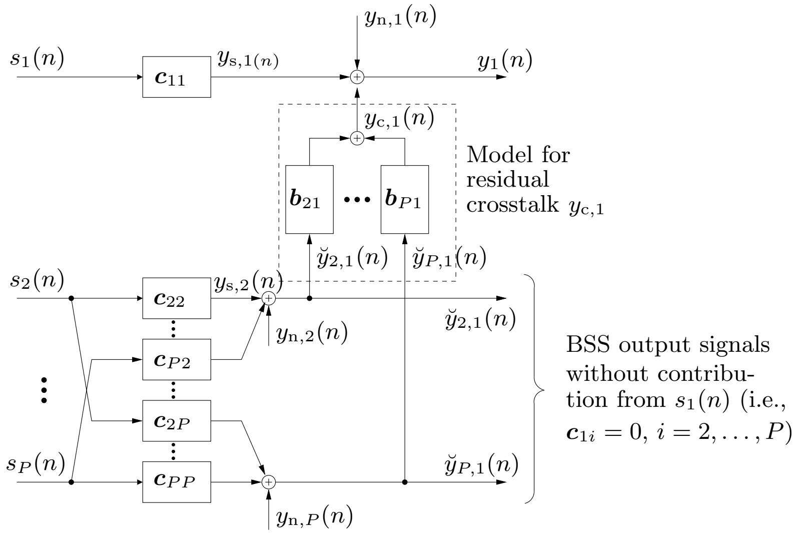


Fig. 1.13. Model of the residual crosstalk component $y_{\mathrm{c},q}$ contained in the q-th BSS output channel y_q illustrated for the first channel, i.e., $q = 1$. In contrast to Fig. 1.12(b) this model is solely based on observable quantities.

filter $\boldsymbol{b}_{21}$ in the least-squares sense [66]. We will see in the following that due to the usage of additional a-priori information this model is nevertheless suitable for the estimation of the residual-cross talk PSD.

The model in Fig. 1.12(b) requires the desired source signal component $y_{\mathrm{s},2}(n)$ in the second BSS output. However, in practice it cannot be assumed that the BSS system always achieves perfect source separation. Especially in the initial convergence phase or with moving sources, there is some residual crosstalk remaining in all outputs. Therefore, we have to modify the residual crosstalk model so that only observable quantities are used. Hence, in Fig. 1.13 the desired signal component $y_{\mathrm{s},i}(n)$ for the i-th channel is replaced by the signal $\breve{y}_{i,q}(n)$ which denotes the BSS output signal of the i-th channel but without any interfering crosstalk components from the q-th point source (i.e., desired source $s_q(n)$). This means that the overall filters $\boldsymbol{c}_{qi}$ from the q-th source to the i-th output ($i = 1, \ldots, P$, $i \neq q$) are assumed to be zero. In practice, this condition is fulfilled by an adaptation control which determines time-frequency points where the desired source $s_q(n)$ is inactive. This a-priori information about desired source absence is important for a good estimation of the residual crosstalk PSD and thus, for achieving additional residual crosstalk cancellation. A detailed discussion of the adaptation control will be given in Sec. 1.3.3.2. Due to the bin-wise application of the single-channel postfilter we will in the following formulate the model in the DFT domain. Consequently, the model for the residual crosstalk in the q-th channel based on observable quantities is expressed for the ν-th DFT bin ($\nu = 1, \ldots, R_{\mathrm{post}}$) as

$$\underline{Y}_{\mathrm{c},q}^{(\nu)}(m) = \sum_{i=1,i\neq q}^{P} \underline{\breve{Y}}_{i,q}^{(\nu)}(m)\,\underline{B}_{i,q}^{(\nu)}(m) = \underline{\breve{\boldsymbol{Y}}}_{q}^{(\nu)^{\mathrm{T}}}(m)\,\underline{\boldsymbol{B}}_{q}^{(\nu)}(m), \tag{1.56}$$

where $\underline{\breve{Y}}_{i,q}^{(\nu)}(m)$ and $\underline{B}_{i,q}^{(\nu)}(m)$ are the DFT-domain representations of $\breve{Y}_{i,q}(m)$ and $\boldsymbol{b}_{iq}(m)$, respectively. The variable $\underline{\breve{\boldsymbol{Y}}}_{q}^{(\nu)}(m)$ is the $P-1$ dimensional DFT-domain column vector containing $\underline{\breve{Y}}_{i,q}^{(\nu)}(m)$ for $i = 1, \ldots, P$, $i \neq q$, and $\underline{\boldsymbol{B}}_{q}^{(\nu)}(m)$ is the column vector containing the unknown filter weights $\underline{B}_{i,q}^{(\nu)}(m)$ for $i = 1, \ldots, P$, $i \neq q$.

Diffuse sound field

It should be pointed out that the adaptation control only ensures that the desired source $s_q(n)$ is absent in the i-th BSS output channel $\underline{\breve{Y}}_{i,q}^{(\nu)}(m)$. However, the background noise $\underline{Y}_{\mathrm{n},i}^{(\nu)}(m)$ is still present in the i-th BSS output channel as can also be seen in Fig. 1.13. If the background noise is spatially correlated between the q-th and i-th BSS output channel, then the coefficient $\underline{B}_{i,q}^{(\nu)}(m)$ would not only model the leakage from the separated source in the i-th channel, but $\underline{B}_{i,q}^{(\nu)}(m)$ would also be affected by the spatially correlated background noise. However, as an additional measure, the noise PSD $\mathrm{E}\{|\underline{Y}_{\mathrm{n},q}^{(\nu)}(m)|^2\}$ is estimated individually in each channel by one of the noise estimation methods known from single-channel speech enhancement. Therefore, if the background noise is already included in the residual crosstalk model, this would lead to an overestimation of the noise PSD. In Sec. 1.3.1 the character of the background noise such as car or babble noise was examined and the correlation of the noise sources between the sensors was evaluated using the magnitude squared coherence (MSC). It was concluded that the MSC of such background noise exhibits the same characteristics as a diffuse sound field leading to strong spatial correlation for low frequencies but to very small spatial correlation at higher frequencies. The model of the residual crosstalk is based on the BSS output signals and hence, it is of interest how the BSS system changes the MSC of the noise signals. In Fig. 1.14(a) and Fig. 1.14(b) the MSC of car noise and babble noise, which was estimated recursively according to Eq. 1.39 with the parameters $R = 512$, $\gamma = 0.9$ is plotted. For the car scenario a two-microphone array with a spacing of 4 cm was mounted at the interior mirror and the driver and co-driver were speaking simultaneously. Then the block-online BSS algorithm given in Eq. 1.35 together with the narrowband normalization [7] was applied. The same experiment was performed with two sources in a reverberant room where the babble noise was generated by a circular loudspeaker array using 16 individual speech signals. As pointed out in the beginning, the BSS algorithm tries to achieve source separation by aiming at mutual independence of the BSS output signals. From Fig. 1.14 it can be seen that in the presence of background noise this also leads to a *spatial decorrelation of the noise signals* at the BSS outputs. The car noise

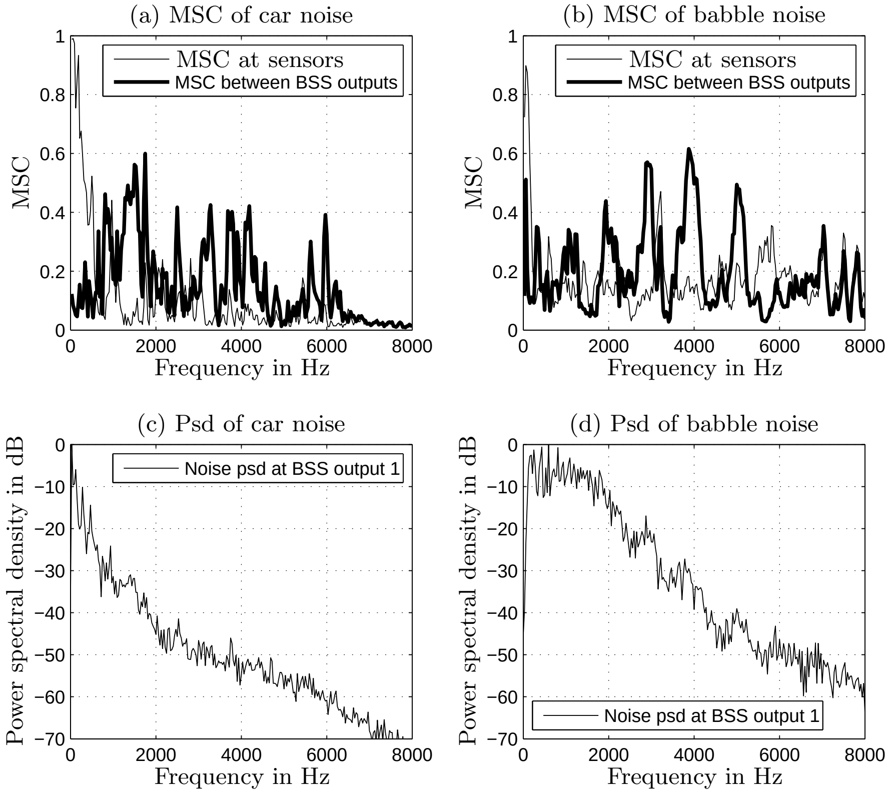


Fig. 1.14. Magnitude-squared coherence (MSC) of the car (a) and speech babble (b) background noise between the sensors and between the BSS outputs. The long-term noise PSDs of car noise and speech babble are shown in (c) and (d), respectively.

which is dominant at low frequencies (see Fig. 1.14(c)) has a MSC close to zero at these frequencies (see Fig. 1.14(a)). Only at higher frequencies, where the noise signal has much less energy, a larger MSC can be observed. The reduction of the MSC for the relevant frequencies can analogously also be observed for the babble noise. This observation shows that the background noise is spatially decorrelated at the BSS outputs and thus, confirms that the model for the residual crosstalk introduced in Eq. 1.56 is valid even in the case of background noise. This also justifies the independent estimations of the background noise in each channel and thus, we can apply noise estimation methods previously derived for single-channel speech enhancement algorithms. The residual cross-talk, however, is correlated across the output channels. These characteristics of residual cross-talk and background noise will be exploited in the next section to derive suitable estimation procedures.

After introducing the residual crosstalk model and validating it for the case of existing background noise, we briefly discuss the relationships to the models used in previous publications on post-processing for BSS. In [72] a Wiener-based approach for residual crosstalk cancellation is presented for the case $P = 2$. There, a greatly simplified model is used where all coefficients $\underline{B}_{i,q}^{(\nu)}(m)$ $(i, q = 1, 2,\ i \neq q)$ are assumed to be equal to one. Similarly, in [80] one constant factor was chosen for all $\underline{B}_{i,q}^{(\nu)}(m)$. A model closer to Eq. 1.56, but based on magnitude spectra, was given in [22] which was then used for the implementation of a spectral subtraction rule. In contrast to the estimation method presented in the next section, the frequency-dependent coefficients of the model were learned by a modified least-mean-squares (LMS) algorithm. In [68] and [69] more sophisticated models were proposed allowing for time-delays or FIR filtering in each DFT bin. The model parameters were estimated by exploiting correlations between the channels or by using an NLMS algorithm. In all of these single-channel approaches the information of the multiple channels is only exploited to estimate the PSDs necessary for the spectral weighting rule. Alternatively, if also the phase of $\underline{Y}_{\mathrm{c},q}^{(\nu)}(m)$ is estimated, then it is also possible to directly subtract the estimate of the crosstalk component $\underline{Y}_{\mathrm{c},q}^{(\nu)}(m)$ from the q-th channel. This was proposed in [57] resulting in an adaptive noise canceller (ANC) structure [86]. The ANC was adapted by a leaky LMS algorithm [37] which includes a variable step size to allow also for strong desired signal activity without the necessity of an adaptation control.

The background noise component in the q-th channel $\underline{Y}_{\mathrm{n},q}^{(\nu)}(m)$ is usually assumed to be more stationary than the desired signal component $\underline{Y}_{\mathrm{s},q}^{(\nu)}(m)$. This assumption is necessary for the noise estimation methods known from single-channel speech enhancement which will be used to estimate the noise PSD $\hat{\mathrm{E}}\{|\underline{Y}_{\mathrm{n},q}^{(\nu)}(m)|^2\}$ in each channel and which are briefly discussed in the next section.

Estimation of Residual Crosstalk and Background Noise Power Spectral Densities

After introducing the residual crosstalk model (Eq. 1.56) we need to estimate the PSDs $\mathrm{E}\{|\underline{Y}_{\mathrm{c},q}^{(\nu)}(m)|^2\}$ of the residual crosstalk and $\mathrm{E}\{|\underline{Y}_{\mathrm{n},q}^{(\nu)}(m)|^2\}$ of the background noise for evaluating Eq. 1.55. To obtain an estimation procedure based on observable quantities we first calculate the cross-power spectral density vector $\underline{\boldsymbol{S}}_{\breve{\boldsymbol{Y}}_q Y_{\mathrm{c},q}}^{(\nu)}(m)$ between $\underline{\breve{\boldsymbol{Y}}}_q^{(\nu)}(m)$ and $\underline{Y}_{\mathrm{c},q}^{(\nu)}(m)$ given as

$$\begin{aligned}\underline{\boldsymbol{S}}_{\breve{\boldsymbol{Y}}_q Y_{\mathrm{c},q}}^{(\nu)}(m) &= \hat{\mathrm{E}}\Big\{\underline{\breve{\boldsymbol{Y}}}_q^{(\nu)^*}(m)\,\underline{Y}_{\mathrm{c},q}^{(\nu)}(m)\Big\}\\ &= \hat{\mathrm{E}}\Big\{\underline{\breve{\boldsymbol{Y}}}_q^{(\nu)^*}(m)\,\underline{\breve{\boldsymbol{Y}}}_q^{(\nu)^{\mathrm{T}}}(m)\Big\}\,\underline{\boldsymbol{B}}_q^{(\nu)}(m)\\ &=: \underline{\boldsymbol{S}}_{\breve{\boldsymbol{Y}}_q \breve{\boldsymbol{Y}}_q}^{(\nu)}(m)\,\underline{\boldsymbol{B}}_q^{(\nu)}(m)\,, \qquad (1.57)\end{aligned}$$

where in the first step $\underline{\boldsymbol{B}}_q^{(\nu)}(m)$ was assumed to be slowly time-varying. Using Eq. 1.56, the power spectral density estimate $\hat{\mathrm{E}}\{|\underline{Y}_{\mathrm{c},q}^{(\nu)}(m)|^2\}$ can be expressed as

$$\begin{aligned}\hat{\mathrm{E}}\Big\{\big|\underline{Y}_{\mathrm{c},q}^{(\nu)}(m)\big|^2\Big\} &= \hat{\mathrm{E}}\Big\{\underline{Y}_{\mathrm{c},q}^{(\nu)^{\mathrm{H}}}(m)\,\underline{Y}_{\mathrm{c},q}^{(\nu)}(m)\Big\} \\ &= \underline{\boldsymbol{B}}_q^{(\nu)^{\mathrm{H}}}(m)\,\underline{\boldsymbol{S}}_{\breve{\boldsymbol{Y}}_q\breve{\boldsymbol{Y}}_q}^{(\nu)}(m)\,\underline{\boldsymbol{B}}_q^{(\nu)}(m)\,. \qquad (1.58)\end{aligned}$$

Solving Eq. 1.57 for $\underline{\boldsymbol{B}}_q^{(\nu)}(m)$ and inserting it into Eq. 1.58 leads to

$$\hat{\mathrm{E}}\Big\{\big|\underline{Y}_{\mathrm{c},q}^{(\nu)}(m)\big|^2\Big\} = \underline{\boldsymbol{S}}_{\breve{\boldsymbol{Y}}_q Y_{\mathrm{c},q}}^{(\nu)^{\mathrm{H}}}(m)\left(\underline{\boldsymbol{S}}_{\breve{\boldsymbol{Y}}_q\breve{\boldsymbol{Y}}_q}^{(\nu)}(m)\right)^{-1}\underline{\boldsymbol{S}}_{\breve{\boldsymbol{Y}}_q Y_{\mathrm{c},q}}^{(\nu)}(m)\,. \qquad (1.59)$$

As $\underline{Y}_{\mathrm{c},q}^{(\nu)}(m)$, $\underline{Y}_{\mathrm{s},q}^{(\nu)}(m)$, and $\underline{Y}_{\mathrm{n},q}^{(\nu)}(m)$ in Fig. 1.13 are assumed to be mutually uncorrelated, $\underline{\boldsymbol{S}}_{\breve{\boldsymbol{Y}}_q Y_{\mathrm{c},q}}^{(\nu)}(m)$ can also be estimated as the cross-power spectral density $\underline{\boldsymbol{S}}_{\breve{\boldsymbol{Y}}_q Y_q}^{(\nu)}(m)$ between $\underline{\breve{\boldsymbol{Y}}}_q^{(\nu)}(m)$ and q-th output of the BSS system $\underline{Y}_q^{(\nu)}(m)$ leading to the final estimation procedure:

$$\hat{\mathrm{E}}\Big\{\big|\underline{Y}_{\mathrm{c},q}^{(\nu)}(m)\big|^2\Big\} = \underline{\boldsymbol{S}}_{\breve{\boldsymbol{Y}}_q Y_q}^{(\nu)^{\mathrm{H}}}(m)\left(\underline{\boldsymbol{S}}_{\breve{\boldsymbol{Y}}_q\breve{\boldsymbol{Y}}_q}^{(\nu)}(m)\right)^{-1}\underline{\boldsymbol{S}}_{\breve{\boldsymbol{Y}}_q Y_q}^{(\nu)}(m). \qquad (1.60)$$

Thus, the power spectral density of the residual crosstalk for the q-th channel can be efficiently estimated in each DFT bin $\nu = 0, \ldots, R-1$ by computing the $1 \times P-1$ cross-power spectral density vector $\underline{\boldsymbol{S}}_{\breve{\boldsymbol{Y}}_q Y_q}^{(\nu)}(m)$ between input and output of the model shown in Fig. 1.13 and calculating the $P-1 \times P-1$ cross-power spectral density matrix $\underline{\boldsymbol{S}}_{\breve{\boldsymbol{Y}}_q\breve{\boldsymbol{Y}}_q}^{(\nu)}(m)$ of the inputs. One possible implementation for estimating this expectation is given by an exponentially weighted average

$$\hat{\mathrm{E}}\big\{a(m)\big\} = (1-\gamma)\sum_i \gamma^{m-i} a(i)\,, \qquad (1.61)$$

where $a(m)$ is the quantity to be averaged. The advantage is that this can also be formulated recursively leading to

$$\underline{\boldsymbol{S}}_{\breve{\boldsymbol{Y}}_q\breve{\boldsymbol{Y}}_q}^{(\nu)}(m) = \gamma\,\underline{\boldsymbol{S}}_{\breve{\boldsymbol{Y}}_q\breve{\boldsymbol{Y}}_q}^{(\nu)}(m-1) + (1-\gamma)\,\underline{\breve{\boldsymbol{Y}}}_q^{(\nu)^*}(m)\,\underline{\breve{\boldsymbol{Y}}}_q^{(\nu)^{\mathrm{T}}}(m)\,, \qquad (1.62)$$

$$\underline{\boldsymbol{S}}_{\breve{\boldsymbol{Y}}_q Y_q}^{(\nu)}(m) = \gamma\,\underline{\boldsymbol{S}}_{\breve{\boldsymbol{Y}}_q Y_q}^{(\nu)}(m-1) + (1-\gamma)\,\underline{\breve{\boldsymbol{Y}}}_q^{(\nu)^*}(m)\,\underline{Y}_q^{(\nu)}(m)\,. \qquad (1.63)$$

In summary, the power spectral density of the residual crosstalk for the q-th channel can be efficiently estimated in each DFT bin $\nu = 0, \ldots, R-1$ using Eq. 1.60 together with the recursive calculation of the $P-1 \times P-1$ cross-power spectral density matrix (Eq. 1.62) and the $P-1\times 1$ cross-power spectral density vector (Eq. 1.63). It should be noted that similar estimation techniques have been used to determine a post-filter for residual echo suppression in the context of acoustic echo cancellation (AEC) [32, 79]. However, the methods

Minimum statistics
Adaptation control

presented in [32, 79] are different in two ways: Firstly, in contrast to BSS where several interfering point sources may be active, the AEC post-filter was derived for a single channel, i.e., the residual echo originates from only one point source and thus all quantities in Eq. 1.60 reduce to scalar values. Secondly, in the AEC problem a reference signal for the echo is available. In BSS however, $\underline{\breve{\boldsymbol{Y}}}_q^{(\nu)}(m)$ is not immediately available as it can only be estimated if the desired source signal in the q-th channel is currently inactive. Strategies how to determine such time intervals are discussed in the next section.

The estimation of the PSD of the background noise $\hat{\mathrm{E}}\{|\underline{Y}_{\mathrm{n},q}^{(\nu)}(m)|^2\}$ is already a long-standing research topic in single-channel speech enhancement and an overview on the various methods can be found, e.g., in [39]. Usually it is assumed that the noise PSD is at least more stationary than the desired speech PSD. The noise estimation can be performed during speech pauses, which have to be detected properly by a voice activity detector. As voice activity detection algorithms are rather unreliable in low SNR conditions, several methods have been proposed which can track the noise PSD continuously. One of the most prominent methods is the minimum statistics approach which is based on the observation that the power of a noisy speech signal frequently decays to the power of the background noise. Hence, by tracking the minima the power spectral density of the noise is obtained. In [63] a minima tracking algorithm was proposed which includes an optimal smoothing of the noise PSD together with a bias correction and which will be applied in the experiments in Sec. 1.3.3.3. An overview of other methods providing continuous noise PSD estimates can be found, e.g., in [39].

Adaptation Control Based on SIR Estimation

In the previous sections it was shown that the estimation of the residual crosstalk power spectral density in the q-th channel is only possible at time instants when the desired point source of the q-th channel is inactive. As pointed out already above, speech signals can be assumed to be sufficiently sparse in the time-frequency domain so that even in reverberant environments regions can be found where one or more sources are inactive (see, e.g., [90]). This fact will be exploited by constructing a DFT-based adaptation control necessary for the estimation of the residual cross-talk PSD. In this section we will first briefly review an adaptation control approach which is already known from the literature on post-processing for BSS. Due to the similarity of the adaptation control necessary for estimating the residual crosstalk and the control necessary for adaptive beamformers applied to acoustic signals, also the existing approaches in the beamforming literature will be briefly summarized. A sophisticated bin-wise adaptation control proposed in [42] in the context of adaptive beamforming will then be applied in a slightly modified version to the post-processing scheme.

In general, all adaptation controls aim at estimating the SIR in the time domain or in a bin-wise fashion in the DFT domain. For the latter, the SIR estimate is given for the ν-th DFT bin as the ratio of the desired signal PSD and the PSD of the interfering signals. Thus, the SIR estimate at the q-th BSS output is given as

$$\widehat{\underline{SIR}}_q^{(\nu)}(m) = \frac{\hat{\mathrm{E}}\Big\{\big|\underline{Y}_{\mathrm{s},q}^{(\nu)}(m)\big|^2\Big\}}{\hat{\mathrm{E}}\Big\{\big|\underline{Y}_{\mathrm{c},q}^{(\nu)}(m)\big|^2\Big\}} \,. \tag{1.64}$$

For the case of a BSS system with two output channels ($P = 2$) together with the assumption that for the number of simultaneously active point sources $Q \leq P$ holds, a simple SIR estimate is given by approximating the desired signal component $\underline{Y}_{\mathrm{s},q}^{(\nu)}(m)$ with the BSS output signal $\underline{Y}_q^{(\nu)}(m)$ of the q-th channel and approximating the interfering signal component by the BSS output signal of the other channel. This yields, e.g., for the approximated SIR estimate in the first BSS output channel

$$\widehat{\underline{SIR}}_1^{(\nu)}(m) \approx \frac{\hat{\mathrm{E}}\Big\{\big|\underline{Y}_1^{(\nu)}(m)\big|^2\Big\}}{\hat{\mathrm{E}}\Big\{\big|\underline{Y}_2^{(\nu)}(m)\big|^2\Big\}} \,. \tag{1.65}$$

This approximation is justified if the BSS system already provides enough separation performance so that the BSS output signals can be seen as estimates of the point sources. In [68, 69] the time-average $\hat{\mathrm{E}}\{\cdot\}$ in Eq. 1.65 has been approximated by taking the instantaneous PSD values and the resulting approximated SIR was used successfully as a decision variable for controlling the estimation of the residual crosstalk. If $\widehat{\underline{SIR}}_1^{(\nu)}(m) < 1$, then the crosstalk $\underline{Y}_{\mathrm{c},1}^{(\nu)}(m)$ was estimated and for $\widehat{\underline{SIR}}_1^{(\nu)}(m) > 1$ the crosstalk of the second channel $\underline{Y}_{\mathrm{c},2}^{(\nu)}(m)$ was determined. In [5] this adaptation control was refined by the introduction of a safety margin Υ to improve reliability. By comparing Eq. 1.65 to a fixed threshold Υ it is ensured that a certain SIR value $\widehat{\underline{SIR}}_1^{(\nu)}(m) < \Upsilon$ has to be attained to allow the conclusion that the desired signal is absent and thus, allow estimation of the residual crosstalk $\underline{Y}_{\mathrm{c},1}^{(\nu)}(m)$. The safety margin Υ has to be chosen between $0 < \Upsilon \leq 1$ and was set in [5] to $\Upsilon = 0.9$. For an extension of this mechanism to $P, Q > 2$ a suitable approximation for $\hat{\mathrm{E}}\{|\underline{Y}_{\mathrm{c},q}^{(\nu)}(m)|^2\}$ in the SIR estimate (Eq. 1.64) is important. In [5] it was suggested for $P, Q > 2$ to use the maximum PSD of the remaining channels $\hat{\mathrm{E}}\{|\underline{Y}_i^{(\nu)}(m)|^2\}$, $i \neq q$. For increasing P, Q this requires a very careful choice of Υ. In such scenarios, it is advantageous to replace the fixed threshold Υ by adaptive thresholding. As we will see in the following, such sophisticated adaptation controls were treated in the beamforming literature and will now be applied to the BSS post-processing scheme.

If adaptive beamformers, such as the generalized sidelobe canceller (GSC) (see, e.g., [42]), are applied to acoustic signals, then usually an adaptation control is required for the adaptive filters aiming at interference cancellation. Analogously to the residual crosstalk estimation procedure discussed in Sec. 1.3.3.2, the adaptation of the adaptive interference canceller has to be stalled in the case of a strong desired signal. This analogy allows to apply the approaches in the literature on adaptive beamforming to the post-processing of the BSS output signals. To control the adaptation of beamformers, a correlation-based method was proposed in [36] and recently in a modified form also in [47]. Another approach relies on the comparison of the outputs of a fixed beamformer with the main lobe steered towards the desired source and a complementary beamformer which steers a spatial null towards the desired source [44]. The ratio of the output signal powers, which constitutes an estimate of the SIR is then compared to a threshold to decide if the adaptation should be stopped. As both methods were suggested in the time-domain, this corresponds to a full-band adaptation control, so that in case of a strong desired signal the adaptation is stopped for all DFT bins. It has been pointed out before that speech signals are sparse in the DFT domain and thus, better performance of the adaptation algorithm can be expected when using a bin-wise adaptation control. This was the motivation in [42] for transferring the approach based on two fixed beamformer outputs to the DFT domain leading to a frequency-dependent SIR estimate. Instead of a fixed threshold $\varUpsilon$, additionally an adaptive threshold $\underline{\varUpsilon}_q^{(\nu)}(m)$ for each channel and DFT bin has been proposed leading to a more robust decision. The application of this adaptation control to the estimation of the residual crosstalk, which is required for the post-processing algorithm, will be discussed in the following.

In [42] the estimate $\hat{\mathrm{E}}\{|\underline{Y}_{\mathrm{s},q}^{(\nu)}(m)|^2\}$ of the desired signal required for the SIR estimate (Eq. 1.64) is obtained by a delay-and-sum beamformer. This requires an array of several microphones which should have a spacing that is sufficiently large to allow the suppression of the interfering signals also at low frequencies. Moreover, the positions of the microphones are assumed to be known. This is in contrast to the BSS application where the sensors can be arbitrarily positioned and where there might be only a small number of sensors available (e.g., $P = 2$). Therefore, instead of a fixed beamformer output we will use the q-th BSS output signal PSD $\hat{\mathrm{E}}\{|\underline{Y}_q^{(\nu)}(m)|^2\}$ as an estimate of the desired signal PSD $\hat{\mathrm{E}}\{|\underline{Y}_{\mathrm{s},q}^{(\nu)}(m)|^2\}$.

The estimate of the interfering signal components required for the SIR estimate (Eq. 1.64) are obtained in [42] by a complementary beamformer which places a spatial null towards the desired source. The difference to the procedure in [44] is that this is done in a bin-wise manner. In our application we will use the PSD of a complementary BSS signal $\underline{\bar{Y}}_q^{(\nu)}$ which is obtained analogously to [42] as

$$\hat{\mathrm{E}}\left\{\left|\underline{\bar{Y}}_q^{(\nu)}(m)\right|^2\right\} = \hat{\mathrm{E}}\left\{\left|\underline{X}_q^{(\nu)}(m)\right|^2\right\} - \hat{\mathrm{E}}\left\{\left|\underline{Y}_q^{(\nu)}(m)\right|^2\right\}. \qquad (1.66)$$

Here it is assumed that the filtering due to the BSS demixing system is approximately linear phase and that the BSS output signal and the microphone signal have been properly time-aligned before subtracting their PSD estimates. It should be noted that due to the usage of a broadband BSS algorithm, the permutation at the BSS output signals is not frequency-dependent. Therefore, a possible permutation of the BSS output channels has no effect on the calculation of the complementary BSS signal.

Usually a recursive average is used for the time-average indicated by the operator $\hat{\mathrm{E}}\{\cdot\}$ which leads to the PSD estimates

$$\underline{S}^{(\nu)}_{x_q x_q}(m) = \gamma\, \underline{S}^{(\nu)}_{x_q x_q}(m-1) + (1-\gamma)\left|\underline{X}^{(\nu)}_q(m)\right|^2, \tag{1.67}$$

$$\underline{S}^{(\nu)}_{y_q y_q}(m) = \gamma\, \underline{S}^{(\nu)}_{y_q y_q}(m-1) + (1-\gamma)\left|\underline{Y}^{(\nu)}_q(m)\right|^2, \tag{1.68}$$

necessary for the estimation of the SIR in the q-th BSS output channel. The SIR estimate (Eq. 1.64) can thus be expressed as

$$\widehat{\underline{SIR}}^{(\nu)}_q(m) \approx \frac{\underline{S}^{(\nu)}_{y_q y_q}(m)}{\underline{S}^{(\nu)}_{x_q x_q}(m) - \underline{S}^{(\nu)}_{y_q y_q}(m)}\,. \tag{1.69}$$

The SIR estimate (Eq. 1.69) is then compared to an adaptive threshold $\underline{\Upsilon}^{(\nu)}_q(m)$. If $\widehat{\underline{SIR}}^{(\nu)}_q(m) < \underline{\Upsilon}^{(\nu)}_q(m)$, then the absence of the desired signal in the q-th channel can be assumed. The adaptive threshold is given as the minimum of SIR estimate $\widehat{\underline{SIR}}^{(\nu)}_q(m)$ which is determined for each DFT bin by taking into account the last D_Υ blocks [63]. In practice D_Υ must be large enough to bridge any peak of desired signal activity but short enough to track the nonstationary SIR variations in case of absence of the desired signal. Here, we choose an interval equivalent to a time period of 1.5 sec. Moreover, for small variations

$$\left|\frac{\widehat{\underline{SIR}}^{(\nu)}_q(m) - \underline{\Upsilon}^{(\nu)}_q(m)}{\underline{\Upsilon}^{(\nu)}_q(m)}\right| \le \Delta\Upsilon \tag{1.70}$$

the threshold $\underline{\Upsilon}^{(\nu)}_q(m)$ is updated immediately. In Fig. 1.15 the SIR estimate $\widehat{\underline{SIR}}^{(\nu)}_q$ and the adaptive threshold $\underline{\Upsilon}^{(\nu)}_q$ determined by minimum tracking are illustrated for the DFT bin corresponding to 1 kHz. The results are based on the output signals of the BSS system applied to the car environment. It can be seen that due to the parameter $\Delta\Upsilon = 0.3$ the threshold follows small changes of the SIR estimate immediately. Moreover, it should be pointed out that the SIR estimate in Fig. 1.15 exhibits high positive values due to the good convergence of the BSS algorithm. This is the reason why even in speech pauses of the desired signal, the SIR estimate does rarely exhibit negative SIR values.

In Fig. 1.16 the decision of the adaptation control is illustrated for the first output channel of the BSS system applied to the car environment ($P = Q = 2$). The desired component, residual crosstalk, and background noise component at the first BSS output are depicted in (a)-(c). The decision of the adaptation control is obtained by estimating the SIR according

Minimum statistics

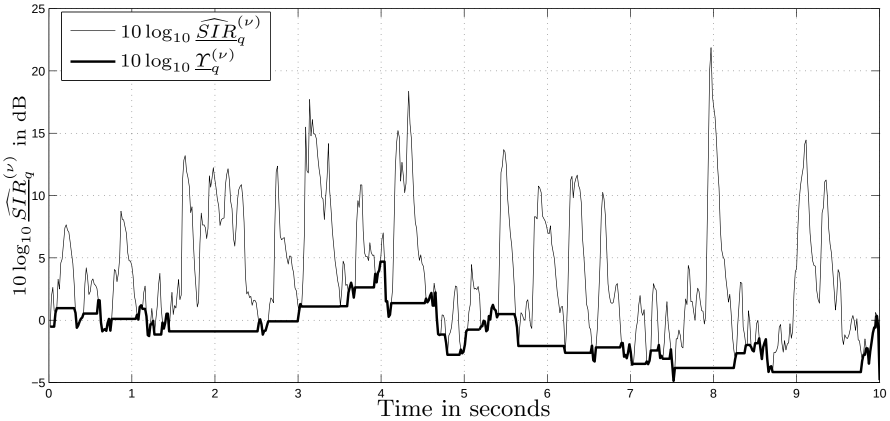


Fig. 1.15. Estimate $10\log_{10}\widehat{\underline{SIR}}_q^{(\nu)}$ of the SIR and adaptive threshold $\underline{\Upsilon}_q^{(\nu)}$ determined by minimum tracking illustrated for the DFT bin corresponding to 1 kHz.

to Eq. 1.69 solely based on observable quantities. Especially due to the existence of background noise $y_{\mathrm{n},q}(n)$ this leads to a biased SIR. Nevertheless, the adaptation control is very robust due to the adaptive threshold $\underline{\Upsilon}_1^{(\nu)}$ based on minimum tracking and the parameter $\Delta\Upsilon = 0.3$ which allows for small variation of the threshold. This can be seen, when comparing the results of the adaptation control with the true SIR illustrated in (e) which is estimated based on unobservable quantities according to Eq. 1.64. In case of high SIR values $\widehat{\underline{SIR}}_1^{(\nu)}(m)$, the desired signal in the first channel is present and the residual crosstalk PSD $\hat{\mathrm{E}}\{|\underline{Y}_{\mathrm{c},2}^{(\nu)}(m)|^2\}$ of the other channel is estimated. Vice versa, a low SIR in the first channel allows to adapt $\hat{\mathrm{E}}\{|\underline{Y}_{\mathrm{c},1}^{(\nu)}(m)|^2\}$.

In case that the adaptation control stalls the estimation of the residual crosstalk for the ν-th DFT bin in one of the P BSS output channels, the residual crosstalk estimate from the previous block has to be used. As speech is a nonstationary process and therefore, the statistics of the residual crosstalk are quickly time-varying, this would deteriorate the performance of the postfilter $\underline{G}_q^{(\nu)}(m)$. On the other hand, as pointed out above, the minimum statistics algorithm can provide continuous noise PSD estimates even in periods with desired signal activity. Therefore, for those time instants where the estimate of residual crosstalk cannot be updated, i.e., where the desired source signal is dominant, a postfilter

$$\underline{G}_{\mathrm{n},q}^{(\nu)}(m) = \frac{\hat{\mathrm{E}}\left\{\left|\underline{Y}_q^{(\nu)}(m)\right|^2\right\} - \hat{\mathrm{E}}\left\{\left|\underline{Y}_{\mathrm{n},q}^{(\nu)}(m)\right|^2\right\}}{\hat{\mathrm{E}}\left\{\left|\underline{Y}_q^{(\nu)}(m)\right|^2\right\}} \tag{1.71}$$

merely aiming at suppression of the background noise is applied.

In Tab. 1.2 the adaptation control and the resulting application of the postfilters is outlined for the q-th BSS output channel.

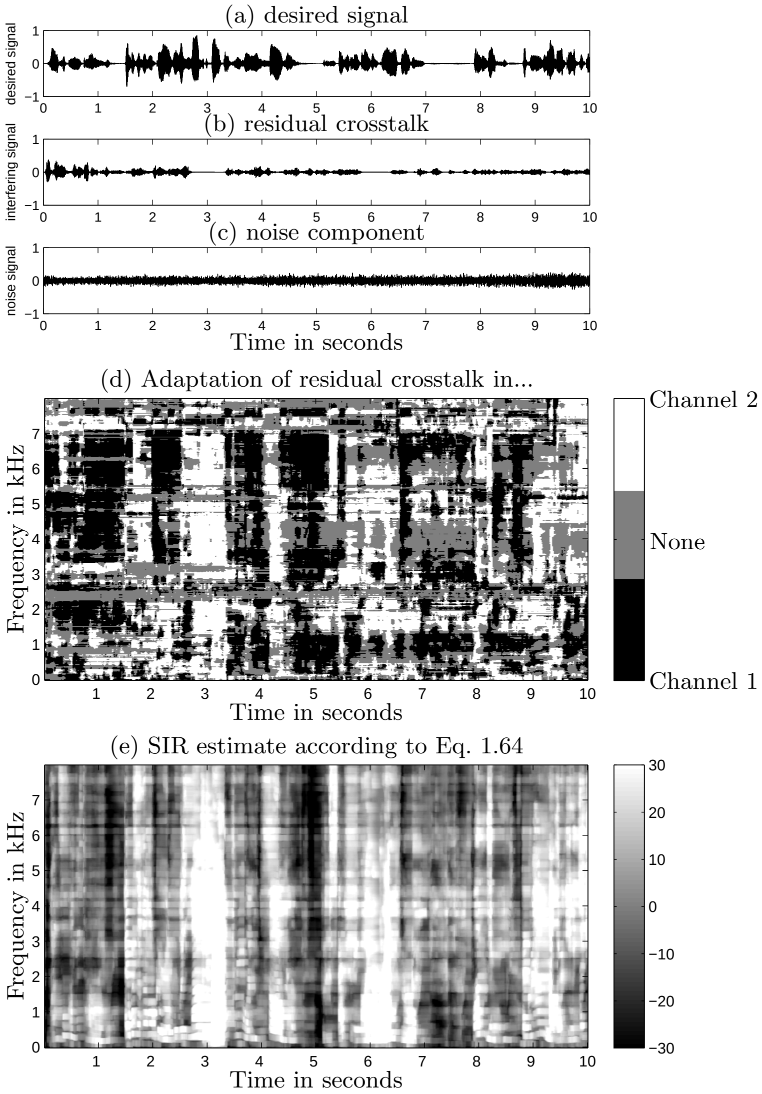


Fig. 1.16. BSS output signal components for the car environment with an input SNR at the sensors of 0 dB showing the desired signal (a), residual crosstalk (b) and background noise (c) in the first channel. Based on the SIR estimate (Eq. 1.69) and the adaptive threshold $\underline{\Upsilon}_1^{(\nu)}(m)$ the decision of the adaptation control is shown in (d). For comparison, the SIR (Eq. 1.64) computed for the true signal components in the first channel is illustrated in (e).

Table 1.2. Adaptation control and application of the postfilter for the q-th BSS output channel and ν-th DFT bin.

Number	Algorithmic part
1.	Estimate $\underline{S}_{x_q x_q}^{(\nu)}(m)$ and $\underline{S}_{y_q y_q}^{(\nu)}(m)$ according to Eqs. 1.67 and 1.68
2.	Estimate $\widehat{\underline{SIR}}_q^{(\nu)}(m)$ according to Eq. 1.69
3.	Estimate $\hat{\mathrm{E}}\{\lvert\underline{Y}_{\mathrm{n},q}^{(\nu)}(m)\rvert^2\}$ by minimum statistics algorithm
4.	Tracking of minima of $\widehat{\underline{SIR}}_q^{(\nu)}(m)$: If $\lvert(\widehat{\underline{SIR}}_q^{(\nu)}(m) - \underline{\Upsilon}_q^{(\nu)}(m))/\underline{\Upsilon}_q^{(\nu)}(m)\rvert \leq \varDelta\Upsilon$ Replace all values of $\underline{\Upsilon}_q^{(\nu)}(i)$ inside the buffer, i.e., $\underline{\Upsilon}_q^{(\nu)}(i) = \widehat{\underline{SIR}}_q^{(\nu)}(m)$, $i = m, \ldots, m - D_\Upsilon + 1$ If $\widehat{\underline{SIR}}_q^{(\nu)}(m)$ is the minimum of $\underline{\Upsilon}_q^{(\nu)}(m-i)$, $i = 0, \ldots, D_\Upsilon - 1$ Set current value of buffer $\underline{\Upsilon}_q^{(\nu)}(m) = \widehat{\underline{SIR}}_q^{(\nu)}(m)$
5.	If minimum is detected, i.e., $\widehat{\underline{SIR}}_q^{(\nu)}(m) \leq \underline{\Upsilon}_q^{(\nu)}(m)$: Calculate residual crosstalk $\hat{\mathrm{E}}\{\lvert\underline{Y}_{\mathrm{c},q}^{(\nu)}(m)\rvert^2\}$ according to Eq. 1.60 Compute postfilter (Eq. 1.55) for residual crosstalk and noise
6.	If no minimum is detected, i.e., $\widehat{\underline{SIR}}_q^{(\nu)}(m) > \underline{\Upsilon}_q^{(\nu)}(m)$: Compute postfilter (Eq. 1.71) for noise only

1.3.3.3 Experimental Results for Reverberant and Noisy Environments

In the evaluation of the postfiltering algorithm summarized in Tab. 1.2 the same two noisy scenarios have been considered as in Sec. 1.3.2 and their description is briefly summarized. The first one is a car environment where a pair of omnidirectional microphones with a spacing of 20 cm was mounted at the interior mirror and recorded a male and female speaker at the driver and co-driver positions, respectively using a sampling rate of $f_\mathrm{s} = 16\,\mathrm{kHz}$. The long-term SNR was adjusted to 0 dB which is a realistic value commonly encountered inside car compartments. The second scenario corresponds to the cocktail party problem which is usually described by the task of listening to one desired point source in the presence of speech babble noise consisting of the utterances of many other speakers. Speech babble is well described by a diffuse sound field, however, there may also be several other distinct noise point sources present. In our experiments we simulated such a cocktail party scenario inside a living room environment where speech babble noise was generated by a circular loudspeaker array with a diameter of 3 m. The two omnidirectional microphones with a spacing of 20 cm were placed in the center of the loudspeaker array from which 16 speech signals were reproduced to simulate the speech babble noise. Additionally, two distinct point sources at a distance of 1 m and at the angles of 0° and −80° were used to simulate the desired and one interfering point source, respectively. The long-term input SNR at the microphones has been adjusted for the living room scenario to 10 dB. This is realistic, as due to the speech-like spectrum of the background noise the microphone signals which exhibit higher SNR values are perceptually already as annoying as those with significantly lower SNR values for lowpass car noise.

The second-order statistics BSS algorithm with the narrowband normalization described in [7] is applied to the two noisy scenarios. To evaluate the performance two measures have been used: The segmental SIR which is defined as the ratio of the signal power of the desired signal to the signal power of the residual crosstalk stemming from point source interferers and the segmental SNR defined as the ratio of the signal power of the desired signal to the signal power of the possibly diffuse background noise. In both cases, the SIR and SNR improvement due to the application of the postfilter is measured and averaged over both channels. The segmental SIR improvement $\mathit{\Delta SIR}_{\mathrm{seg}}(m)$ is plotted as a function of the block index m to illustrate the convergence effect of the BSS system. The channel-averaged segmental SNR improvement $\mathit{\Delta SNR}_{\mathrm{seg}}$ is given as the average over all blocks. To assess the desired signal distortion, the unweighted log-spectral distance (SD) which describes the Euclidean distance between logarithmic short-time magnitude spectra has been measured between the desired signal at the input and the output of the postfilter and is given as

$$SD_{s_r,q} = \frac{1}{K_S}\sum_{m=1}^{K_S}\sqrt{\frac{1}{R}\sum_{\nu=0}^{R-1}\left(20\ \mathrm{lg}\,\frac{\left|\underline{Z}_{\mathrm{s},q}^{(\nu)}(m)\right|}{\left|\underline{Y}_{\mathrm{s},q}^{(\nu)}(m)\right|}\right)^2}\,. \tag{1.72}$$

The DFT length R for computing $SD_{s_r,q}$ is usually set to be small so that speech can be assumed stationary. In our experiments we used $R = 256$ and set K_S large enough to cover the whole signal length. To reduce artifacts such as, e.g., musical noise, the postfilter (Eq. 1.55) is usually calculated using an adaptive oversubtraction factor $\xi_q^{(\nu)}$ as proposed in [12]. Moreover, negative gains of the postfilters are set to zero. Hence in the experiments the postfilter

$$\begin{aligned}&\underline{G}_q^{(\nu)}(m)\\ &= \frac{\max\left\{\left(\hat{\mathrm{E}}\left\{\left|\underline{Y}_q^{(\nu)}(m)\right|^2\right\} - \xi_q^{(\nu)}\left(\hat{\mathrm{E}}\left\{\left|\underline{Y}_{\mathrm{n},q}^{(\nu)}(m)\right|^2\right\} + \hat{\mathrm{E}}\left\{\left|\underline{Y}_{\mathrm{c},q}^{(\nu)}(m)\right|^2\right\}\right)\right), 0\right\}}{\hat{\mathrm{E}}\left\{\left|\underline{Y}_q^{(\nu)}(m)\right|^2\right\}}\end{aligned} \tag{1.73}$$

was used. For the post-processing algorithm, $\gamma = 0.9$ and a DFT length of $R_{\mathrm{post}} = 2048$ was chosen. The block length N_{post} was equal to the DFT length and an overlap factor $\alpha = 4$ was used. The parameters of the adaptation control are given as $\Delta\Upsilon = 0.3$ and $D_{\Upsilon} = 94$ corresponding to a period of 1.5 sec over which the minimum is tracked.

In Fig. 1.17 the results for the separation of the two speech point sources can be seen. For both scenarios the separation performance of the combined system of BSS and single-channel postfilter (solid) outperforms the BSS performance (dashed). In contrast to the BSS system which possesses an inherent adaptation control implied by the normalization term in the update equation, the postfilter relies on a-priori information provided by the adaptation control. Hence, it is possible to accurately estimate the residual crosstalk at the BSS

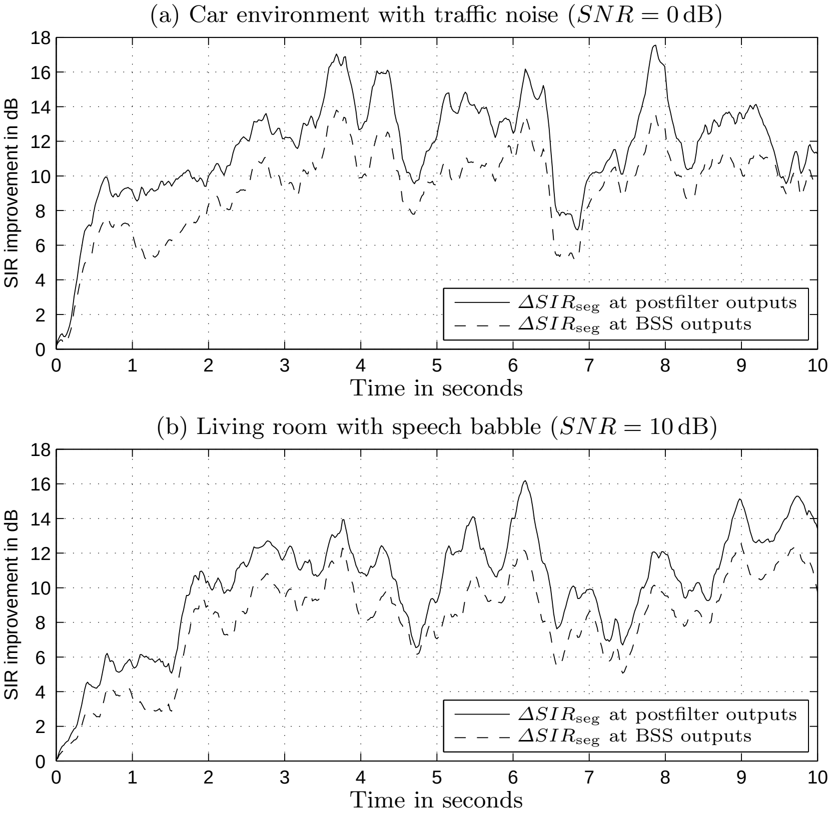


Fig. 1.17. Segmental SIR improvement $\Delta SIR_{\mathrm{seg}}(m)$ depicted over time for two environments containing two speech point-source and additional background noise: (a) car compartment with background noise consisting of car and traffic noise ($SNR = 0\,\mathrm{dB}$) and (b) living room scenario with speech babble background noise from 16 speakers ($SNR = 10\,\mathrm{dB}$). Speech separation results are shown for BSS outputs and postfilter outputs.

Table 1.3. Segmental SNR and unweighted log-spectral distortion for both scenarios

Scenario	ΔSNR_{seg} at BSS outputs	ΔSNR_{seg} at postfilter outputs	SD at postfilter outputs
Car	3.0 dB	4.9 dB	1.0 dB
Cocktail party	0.2 dB	1.3 dB	1.6 dB

outputs for further improvement of the speech separation performance. The reduced absolute level of the SIR improvement in the cocktail party scenario, i.e., in the reverberant living room (Fig. 1.17(b)) is due to longer reverberation and especially due to the background babble noise which exhibits a speech-like long-term spectrum.

Moreover, in both scenarios also the background noise could be partially suppressed. In Tab. 1.3 the segmental SNR averaged over all output channels of the BSS system and of the postfilter is shown. It can be observed that the postfilter achieves an additional SNR gain. As the car noise is more stationary compared to the speech babble noise, the minimum statistics algorithm can better estimate the noise PSD and thus a higher SNR improvement can be achieved by the postfilter.

To assess the speech quality, the SD (Eq. 1.72) between the desired signal at the input and output of the post-filter was calculated and averaged over both output channels. The small values in Tab. 1.3 indicate that the quality of the desired signal is preserved. This was also confirmed by informal listening tests where no musical noise was observed.

1.4 Conclusions

In this chapter we have presented a review of the TRINICON framework which allows to derive BSS algorithms simultaneously exploiting the signal properties nongaussianity, nonwhiteness, and nonstationarity. After the introduction of a generic natural gradient algorithm several special cases leading to efficient implementations have been discussed. It was also outlined how broadband BSS algorithms can be obtained from the TRINICON framework without introducing ambiguities appearing in narrowband algorithms. This has been supported by experimental results in a reverberant room. Subsequently, the application of BSS in noisy environments has been discussed. First, it has been shown that realistic background noise can often be described by the diffuse sound field. As such sound fields have to be modeled by an infinite or at least large number of point sources, the BSS approach only achieves limited noise reduction. Therefore, the extension of the TRINICON framework with pre- and post-processing approaches has been examined. It was shown that a single-channel postfilter applied to each BSS output signal can yield better results than bias-removal techniques used for pre-processing. The postfilter allowed to simultaneously address the cancellation of the residual crosstalk from point source interferers and the suppression of background noise. This was achieved by developing a model for the residual crosstalk and by using a-priori information provided by an adaptation control. The experiments in a car environment and a cocktail-party scenario with background babble noise showed good results for the complemented BSS algorithm. Thus, it can be concluded that by applying the presented post-processing approach, the versatility of the TRINICON BSS algorithms can be extended, resulting in a simultaneous separation of point sources and attenuation of background noise.

1.5 Anmerkungen der Editoren:

- Stimmen in Gl. 1.11 die Indices i von $\beta(i,m)$ und $\hat{p}(...(iL+j))$ zusammen? Muesste $\beta(i,m)$ nicht hinter die zweite Summe wandern und als $\beta(iL+j,m)$ geschrieben werden?

References

1. M. Abramowitz, I.A. Stegun (eds.): *Handbook of Mathematical Functions,* New York, NY, USA: Dover Publications, 1972.
2. R. Aichner, S. Araki, S. Makino, T. Nishikawa, H. Saruwatari: Time-domain blind source separation of non-stationary convolved signals by utilizing geometric beamforming, *Proc. NNSP '02*, 445–454, Martigny, Switzerland, September 2002.
3. R. Aichner, H. Buchner, W. Kellermann: Convolutive blind source separation for noisy mixtures, *Proc. CFA/DAGA '04*, 583–584, Strasbourg, France, March 2004.
4. R. Aichner, H. Buchner, W. Kellermann: On the causality problem in time-domain blind source separation and deconvolution algorithms, *Proc. ICASSP '05*, **5**, 181–184, Philadelphia, PA, USA, March 2005.
5. R. Aichner, M. Zourub, H. Buchner, W. Kellermann: Post-processing for convolutive blind source separation, *Proc. ICASSP '06*, **5**, 37–40, Toulouse, France, May 2006.
6. R. Aichner, H. Buchner, F. Yan, W. Kellermann: A real-time blind source separation scheme and its application to reverberant and noisy acoustic environments, *Signal Processing*, **86**(6), 1260–1277, June 2006.
7. R. Aichner, H. Buchner, W. Kellermann: Exploiting narrowband efficiency for broadband convolutive blind source separation, *EURASIP Journal on Applied Signal Processing*, 1–9, September 2006.
8. R. Aichner: *Acoustic Blind Source Separation in Reverberant and Noisy Environments,* PhD thesis, Universität Erlangen-Nürnberg, Erlangen, Germany, 2007.

9. S.-I. Amari: Natural gradient works efficiently in learning, *Neural Computation*, **10**, 251–276, 1998.
10. S. Araki, R. Mukai, S. Makino, T. Nishikawa, H. Saruwatari: The fundamental limitation of frequency-domain blind source separation for convolutive mixtures of speech, *IEEE Trans. Speech Audio Processing*, **11**(2), 109–116, March 2003.
11. B. Ayad, G. Faucon: Acoustic echo and noise cancelling for hands-free communication systems, *Proc. IWAENC '95*, 91–94, Røros, Norway, June 1995.
12. M. Berouti, R. Schwartz, J. Makhoul: Enhancement of speech corrupted by acoustic noise, *Proc. ICASSP '79*, 208–211, April 1979.
13. H. Brehm, W. Stammler: Description and generation of spherically invariant speech-model signals, *Signal Processing*, **12**, 119–141, 1987.
14. H. Buchner, R. Aichner, W. Kellermann: Blind source separation algorithms for convolutive mixtures exploiting nongaussianity, nonwhiteness, and nonstationarity, *Proc. IWAENC '03*, 275–278, Kyoto, Japan, September 2003.
15. H. Buchner, R. Aichner, W. Kellermann: TRINICON: A versatile framework for multichannel blind signal processing, *Proc. ICASSP' 04*, **3**, 889–892, Montreal, Canada, May 2004.
16. H. Buchner, R. Aichner, W. Kellermann: Blind source separation for convolutive mixtures: A unified treatment, in J. Benesty, Y. Huang (eds.), *Audio Signal Processing for Next-Generation Multimedia Communication Systems*, 255–293, Boston, MA, USA: Kluwer, 2004.
17. H. Buchner, R. Aichner, W. Kellermann: A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics, *IEEE Trans. Speech Audio Processing*, **13**(1), 120–134, January 2005.
18. H. Buchner, R. Aichner, J. Stenglein, H. Teutsch, W. Kellermann: Simultaneous localization of multiple sound sources using blind adaptive MIMO filtering, *Proc. ICASSP '05*, **3**, 97–100, Philadelphia, PA, USA, March 2005.
19. H. Buchner, J. Benesty, W. Kellermann: Generalized multichannel frequency-domain adaptive filtering: Efficient realization and application to hands-free speech communication, *Signal Processing*, **85**, 549–570, 2005.
20. H. Buchner, R. Aichner, W. Kellermann: TRINICON-based blind system identification with application to multiple-source localization and separation, in S. Makino, T.-W. Lee, S. Sawada (eds.), *Blind Speech Separation*, Berlin, Germany: Springer, 2007.
21. J.-F. Cardoso, B.H. Laheld: Equivariant adaptive source separation, *IEEE Trans. Signal Processing*, **44**(12), 3017–3030, December 1996.
22. C. Choi, G.-J. Jang, Y. Lee, S. R. Kim: Adaptive cross-channel interference cancellation on blind source separation outputs, *Proc. ICA '04*, 857–864, Granada, Spain, September 2004.
23. A. Cichocki, R. Unbehauen: *Neural Networks for Optimization and Signal Processing,* Chichester, USA: Wiley, 1994.
24. A. Cichocki, S. Douglas, S.-I. Amari: Robust techniques for independent component analysis (ICA) with noisy data, *Neurocomputing*, **22**, 113–129, 1998.
25. A. Cichocki, S.-I. Amari: *Adaptive Blind Signal and Image Processing,* Chichester, USA: Wiley, 2002.
26. R. K. Cook, R. V. Waterhouse, R. D. Berendt, S. Edelman, M.C. Thompson, Jr.: Measurement of correlation coefficients in reverberant sound fields, *JASA*, **27**(6), 1072–1077, November 1955.
27. T. M. Cover, J. A. Thomas: *Elements of Information Theory,* New York, NY, USA: Wiley, 1991.

28. W. B. Davenport: An experimental study of speech wave propability distribution, *JASA*, **24**(4), 390–399, 1952.
29. P. Divenyi (ed.): *Speech Separation by Humans and Machines,* Norwell, MA, USA: Kluwer, 2005.
30. S. C. Douglas, A. Cichocki, S.-I. Amari: A bias removal technique for blind source separation with noisy measurements, *Electronic Letters*, **34**(14), 1379–1380, July 1998.
31. T. Eltoft, T. Kim, T.-W. Lee: On the multivariate Laplace distribution, *IEEE Signal Processing Lett.*, **13**(5), 300–303, May 2006.
32. G. Enzner, R. Martin, P. Vary: Partitioned residual echo power estimation for frequency-domain acoustic echo cancellation and postfiltering, *Eur. Trans. Telecommun.*, **13**(2), 103–114, 2002.
33. C. L. Fancourt, L. Parra: The coherence function in blind source separation of convolutive mixtures of non-stationary signals, *Proc. NNSP '01*, 303–312, 2001.
34. S. Gazor, W. Zhang: Speech propability distribution, *IEEE Signal Processing Lett.*, **10**(7), 204–207, July 2003.
35. J. Goldman: Detection in the presence of spherically symmetric random vectors, *IEEE Trans. Inform. Theory*, **22**(1), 52–59, January 1976.
36. J. E. Greenberg, P. M. Zurek: Evaluation of an adaptive beamforming method for hearing aids, *JASA*, **91**(3), 1662–1676, March 1992.
37. J. E. Greenberg: Modified LMS algorithms for speech processing with an adaptive noise canceller, *IEEE Trans. Speech Audio Processing*, **6**(4), 338–351, 1998.
38. D. W. Griffin, J. S. Lim: Signal estimation from modified short-time fourier transform, *IEEE Trans. Acoust., Speech, Signal Processing*, **ASSP-32**(2), 236–243, April 1984.
39. E. Hänsler, G. Schmidt: *Acoustic Echo and Noise Control: A Practical Approach,* Hoboken, NJ, USA: Wiley, 2004.
40. D. A. Harville: *Matrix Algebra from a Statistician's Perspective,* Berlin, Germany: Springer, 1997.
41. S. Haykin: *Adaptive Filter Theory,* 4th ed., Englewood Cliffs, NJ, USA: Prentice-Hall, 2002.
42. W. Herbordt: *Sound Capture for Human/Machine Interfaces – Practical Aspects of Microphone Array Signal Processing*, volume 315 of *Lecture Notes in Control and Information Sciences*, Berlin, Germany: Springer, 2005.
43. A. Hiroe: Solution of permutation problem in frequency domain ICA, using multivariate probability density functions. *Proc. ICA '06*, 601–608, Charleston, SC, USA, March 2006.
44. O. Hoshuyama, A. Sugiyama: An adaptive microphone array with good sound quality using auxiliary fixed beamformers and its DSP implementation, *Proc. ICASSP '99*, 949–952, Phoenix, AZ, USA, March 1999.
45. R. Hu, Y. Zhao: Adaptive decorrelation filtering algorithm for speech source separation in uncorrelated noises, *Proc. ICASSP '05*, **1**, 1113–1115, Philadelphia, PA, USA, May 2005.
46. R. Hu, Y. Zhao: Fast noise compensation for speech separation in diffuse noise, *Proc. ICASSP '06*, **5**, 865–868, Toulouse, France, May 2006.
47. T. P. Hua, A. Sugiyama, R. Le Bouquin Jeannes, G. Faucon: Estimation of the signal-to-interference ratio based on normalized cross-correlation with symmetric leaky blocking matrices in adaptive microphone arrays, *Proc. IWAENC '06*, 1–4, Paris, France, September 2006.

48. A. Hyvaerinen, J. Karhunen, E. Oja: *Independent Component Analysis,* New York, NY, USA: Wiley, 2001.
49. S. Ikeda, N. Murata: A method of ICA in time-frequency-domain, *Proc. ICA '99*, 365–371, January 1999.
50. M. Kawamoto, K. Matsuoka, N. Ohnishi: A method of blind separation for convolved non-stationary signals, *Neurocomputing*, **22**, 157–171, 1998.
51. W. Kellermann, H. Buchner, R. Aichner: Separating convolutive mixtures with TRINICON, *Proc. ICASSP '06*, **5**, 961–964, Toulouse, France, May 2006.
52. T. Kim, T. Eltoft, T.-W. Lee: Independent vector analysis: An extension of ICA to multivariate components, *Proc.ICA '06*, 175–172, Charleston, SC, USA, March 2006.
53. S. Kotz, T. Kozubowski, K. Podgorski: *The Laplace Distribution and Generalizations,* Basel, Switzerland: Birkhäuser Verlag, 2001.
54. B. S. Krongold, D.L. Jones: Blind source separation of nonstationary convolutively mixed signals, *Proc. SSAP '00*, 53–57, Pocono Manor, PA, USA, August 2000.
55. S. Kurita, H. Saruwatari, S. Kajita, K. Takeda, F. Itakura: Evaluation of blind signal separation method using directivity pattern under reverberant conditions, *Proc. ICASSP '00*, **5**, 3140–3143, Istanbul, Turkey, June 2000.
56. H. Kuttruff: *Room Acoustics,* 4th ed., London, GB: Spon Press, 2000.
57. S. Y. Low, S. Nordholm, R. Tognieri: Convolutive blind signal separation with post-processing, *IEEE Trans. Speech Audio Processing*, **12**(5), 539–548, September 2004.
58. J. D. Markel, A. H. Gray: *Linear Prediction of Speech,* Berlin, Germany: Springer, 1976.
59. R. Martin, J. Altenhöner: Coupled adaptive filters for acoustic echo control and noise reduction, *Proc. ICASSP '95*, 3043–3046, Detroit, MI, USA, May 1995.
60. R. Martin: *Freisprecheinrichtungen mit mehrkanaliger Echokompensation und Störgeräuschreduktion,* PhD thesis, RWTH Aachen, Aachen, Germany, June 1995 (in German).
61. R. Martin: The echo shaping approach to acoustic echo control, *Speech Communication*, **20**, 181–190, 1996.
62. R. Martin: Small microphone arrays with postfilters for noise and acoustic echo reduction, in M. Brandstein, D. Ward (eds.), *Microphone Arrays: Signal Processing Techniques and Applications*, 255–279, Berlin, Germany: Springer, 2001.
63. R. Martin: Noise power spectral density estimation based on optimal smoothing and minimum statistics, *IEEE Trans. Speech Audio Processing*, **9**(5), 504–512, July 2001.
64. K. Matsuoka, M. Ohya, M. Kawamoto: Neural net for blind separation of nonstationary signals, *IEEE Trans. Neural Networks*, **8**(3), 411–419, 1995.
65. K. Matsuoka, S. Nakashima: Minimal distortion principle for blind source separation, *Proc. ICA '01*, 722–727, San Diego, CA, USA, December 2001.
66. M. Miyoshi, Y. Kaneda: Inverse filtering of room acoustics, *IEEE Trans. Acoust., Speech, Signal Processing*, **36**(2), 145–152, February 1988.
67. L. Molgedey, H. G. Schuster: Separation of a mixture of independent signals using time delayed correlations, *Physical Review Letters*, **72**, 3634–3636, 1994.
68. R. Mukai, S. Araki, H. Sawada, S. Makino: Removal of residual cross-talk components in blind source separation using time-delayed spectral subtraction, *Proc. ICASSP '02*, **2**, 1789–1792, Orlando, FL, USA, May 2002.

69. R. Mukai, S. Araki, H. Sawada, S. Makino: Removal of residual cross-talk components in blind source separation using LMS filters, *Proc. NNSP '02*, 435–444, Martigny, Switzerland, September 2002.
70. T. Nishikawa, H. Saruwatari, K. Shikano: Comparison of time-domain ICA, frequency-domain ICA and multistage ICA for blind source separation, *Proc. EUSIPCO 03*, **2**, 15–18, September 2002.
71. A. Papoulis: *Probability, Random Variables, and Stochastic Processes,* 4th ed., Boston, MA, USA: McGraw-Hill, 2002.
72. K. S. Park, J. S. Park, K. S. Son, H. T. Kim: Postprocessing with Wiener filtering technique for reducing residual crosstalk in blind source separation, *IEEE Signal Processing Lett.*, **13**(12), 749–751, December 2006.
73. L. Parra, C. Spence: Convolutive blind source separation of non-stationary sources, *IEEE Trans. Speech Audio Processing*, **8**(3), 320–327, May 2000.
74. L. Parra, C. Spence, P. Sajda: Higher-order statistical properties arising from the non-stationarity of natural signals, *Advances in Neural Information Processing Systems*, **13**, 786–792, Cambridge, MA, USA: MIT Press, 2000.
75. H. Sawada, R. Mukai, S. de la Kethulle de Ryhove, S. Araki, S. Makino: Spectral smoothing for frequency-domain blind source separation, *Proc. IWAENC '03*, 311–314, Kyoto, Japan, September 2003.
76. K. U. Simmer, J. Bitzer, C. Marro: Post-filtering techniques, in M. Brandstein, D. Ward (eds.), *Microphone Arrays: Signal Processing Techniques and Applications*, 39–60, Berlin, Germany: Springer, 2001.
77. P. Smaragdis: Blind separation of convolved mixtures in the frequency domain, *Neurocomputing*, **22**, 21–34, 1998.
78. L. Tong, R.-W. Liu, V.C. Soon, Y.-F. Huang: Indeterminacy and identifiability of blind identification, *IEEE Trans. on Circuits and Systems*, **38**(5), 499–509, May 1991.
79. V. Turbin, A. Gilloire, P. Scalart, C. Beaugeant: Using psychoacoustic criteria in acoustic echo cancellation algorithms, *Proc. IWAENC '97*, 53–56, London, UK, September 1997.
80. J.-M. Valin, J. Rouat, F. Michaud: Microphone array post-filter for separation of simultaneous non-stationary sources, *Proc. ICASSP '04*, **1**, 221–224, Montreal, Canada, May 2004.
81. S. Van Gerven, D. Van Compernolle: Signal separation by symmetric adaptive decorrelation: Stability, convergence and uniqueness, *IEEE Trans. Signal Processing*, **43**(7), 1602–1612, July 1995.
82. E. Visser, T.-W. Lee: Speech enhancement using blind source separation and two-channel energy based speaker detection, *Proc. ICASSP '03*, **1**, 836–839, HongKong, April 2003.
83. D. Wang and J. Lim: The unimportance of phase in speech enhancement, *IEEE Trans. Acoust., Speech, Signal Processing*, **ASSP-30**(4), 679–681, August 1982.
84. D. Wang, G. J. Brown (eds.): *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications,* New York, NY, USA: Wiley, 2006.
85. E. Weinstein, M. Feder, A. Oppenheim: Multi-channel signal separation by decorrelation, *IEEE Trans. Speech Audio Processing*, **1**(4), 405–413, October 1993.
86. B. Widrow, J. Glover, J. MacCool, J. Kautnitz, C. Williams, R. Hearn, J. Zeidler, E. Dong, R. Goodlin: Adaptive noise cancelling: principles and applications, *Proc. IEEE*, **63**, 1692–1716, 1975.

87. S. Winter, W. Kellermann, H. Sawada, S. Makino: MAP-based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and ℓ_1-norm minimization, *EURASIP Journal on Applied Signal Processing*, 1–12, 2007.

88. H.-C. Wu, J. C. Principe: Simultaneous diagonalization in the frequency domain (SDIF) for source separation, *Proc. ICA '99*, 245–250, Aussois, France, December 1999.

89. K. Yao: A representation theorem and its applications to spherically-invariant random processes, *IEEE Trans. Inform. Theory*, **19**(5), 600–608, September 1973.

90. O. Yilmaz, S. Rickard: Blind separation of speech mixtures via time-frequency masking, *IEEE Trans. Signal Processing*, **52**(7), 1830–1847, July 2004.

Part VI

Selected Applications

Index